Entwicklung bioinformatischer Methoden zur Vorhersage von Sekundärmetabolit-Gen-Clustern in Pilzen

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium **(Dr. rer. nat.)**

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von Diplom-Bioinformatiker Thomas Wolf
geboren am 04.02.1987 in Zwickau

Gutachter

Prof. Dr. Reinhard Guthke, Hans-Knöll-Institut Jena

Prof. Dr. Severin Sasso, Friedrich-Schiller-Universität Jena

Prof. Dr. Edgar Wingender, Georg-August-Universität Göttingen

Verteidigung

12.05.2017, Jena

»Was wir hinterlassen, ist nicht so wichtig, wie die Art, wie wir gelebt haben.«

Jean-Luc Picard

Danksagung

Zu Beginn möchte ich mich bei verschiedenen Personen bedanken, die zum Gelingen der vorliegenden Promotionsarbeit maßgeblich beigetragen haben.

Ursprünglich hatte ich mich am Fritz-Lipmann-Institut (FLI) für die Durchführung meiner Diplomarbeit beworben. Durch einen Zufall bin ich aber am Hans-Knöll-Institut (HKI) in der Forschungsgruppe von Reinhard Guthke gelandet – nur ein paar Meter vom FLI entfernt. Für diesen Zufall bin ich sehr dankbar. Einen menschenfreundlicheren Gruppenleiter und Betreuer als Reinhard Guthke kann man sich, meiner Meinung nach, nicht wünschen.

Mein erster Ansprechpartner am HKI war Ekaterina Shelest. Ob persönlich, per E-Mail oder Telefon – sie hatte immer Zeit für meine Fragen und war stets für intensive wissenschaftliche Diskussionen bereit.

Für die finanzielle Ermöglichung meines Promotionsprojektes danke ich der »International Leibniz Research School«.

Thorsten Heinekamp danke ich dafür, dass er die Überlappungen in den Genomannotationen (Unterabschnitt 3.4.3) untersucht hat.

Vielen Dank auch an Daniel Scharf für die Bereitstellung der Bindestellen des GliZ-Transkriptionsfaktors (Unterabschnitt 3.3.3).

Meinen Kollegen in den beiden Forschungsgruppen »Angewandte Systembiologie« und »Systembiologie / Bioinformatik« des HKI danke ich vor allem für die Kaffeepausen, in denen einige gute Ideen für meine Arbeit entstanden sind. Die Arbeit an der Kaffeewaage hat unglaublich viel Spaß gemacht!

Meinen Kollegen vom FLI danke ich dafür, dass sie mich täglich mit zur Mittagspause genommen haben. Der kurze Spaziergang zur Mensa und die Gespräche waren eine willkommene Abwechslung.

Natürlich danke ich auch allen Korrekturlesern. Eine andere Perspektive aus einem anderen Fachgebiet ist viel Wert. Ein Informatiker zum Beispiel entdeckt sofort, dass die für jeden Biologen und Bioinformatiker verständliche Angabe »1 kb« nicht nur für 1000 Basenpaare stehen kann, sondern auch für die viel gebräuchlicheren Einheiten »Kilobyte« oder »Kilobit« – und deswegen erklärt werden sollte.

Nicht zuletzt gilt mein Dank meiner Familie, meiner Schwiegerfamilie und vor allem meiner geliebten Freundin. Diese Menschen stellen einen wichtigen Teil des eigenen Lebens dar; vor allem wenn man arbeitet um zu leben, statt lebt um zu arbeiten.

Zusammenfassung

Sekundärmetabolite sind eine Gruppe von äußerst vielseitigen Naturstoffen. Sie werden hauptsächlich von Pflanzen, Bakterien und Pilzen produziert, um auf die Veränderung von Umweltbedingungen zu reagieren oder sich gegen andere Organismen zu verteidigen. Viele Sekundärmetabolite sind für den Menschen von hohem medizinischen, pharmazeutischen oder landwirtschaftlichem Interesse. Die Gene für die Biosynthese von Sekundärmetaboliten sind oft in Gen-Clustern organisiert.

Gen-Cluster bestehen aus kolokalisierten und koregulierten Genen, die meistens zum gleichen Stoffwechselweg gehören. Gen-Cluster wurden sowohl in eukaryotischen als auch in prokaryotischen Genomen nachgewiesen, jedoch ist bei vielen Clustern die Funktionweise der beteiligten Gene unbekannt. Sekundärmetabolit-Gen-Cluster sind eine Spezialform der Gen-Cluster. Die Gene dieser Cluster sind in der Nähe von einem oder mehreren Hauptenzymen (»Ankergenen«) angeordnet und besonders eng kolokalisiert. Außerdem gehört zu den Genen im Cluster oft ein cluster-spezifischer Transkriptionsfaktor.

Durch die computergestützte Vorhersage von Ankergenen und Gen-Clustern können deren Regulationsmechanismen und Funktionsweisen besser verstanden werden. Die vorliegende Arbeit beschäftigt sich mit der Entwicklung und Anwendung neuartiger Methoden zur Vorhersage von Sekundärmetabolit-Gen-Clustern und Ankergenen. Im Gegensatz zu bereits verfügbaren Methoden basiert diese neuartige Cluster-Vorhersagemethode insbesondere auf der Hypothese der Koregulation: cluster-spezifische genregulatorische Elemente sollten in den Promotorregionen der Cluster-Gene häufiger vorkommen als an anderen Stellen im Genom. Auf diese Weise können Cluster von Nicht-Cluster-Genen unterschieden werden.

Abstract

Secondary metabolites are structurally diverse natural products. They are mainly deployed by plants, bacteria, and fungi to adapt to changes in the environment and to defend against competing organisms. Many secondary metabolites are also of high medical, pharmaceutical, or agricultural importance. Genes involved in their biosynthesis are often organized in clusters.

Gene clusters are composed of co-localized and co-regulated genes, which often belong to the same metabolic pathway. The clustering of functionally related genes is a common characteristic of both prokaryotic and eukaryotic genomes. However, the structures and biochemical details of many gene clusters are still unknown. Secondary metabolite gene clusters are a special type of gene clusters. The genes of such a cluster are very tightly co-localized. These clusters consist of one or more secondary metabolite key enzymes (»anchor genes«), and in many cases a cluster-specific transcription factor.

The computational prediction of anchor genes and gene clusters, to help understanding their biosynthesis and functions, is a challenging task. This work is about the development and application of novel methods to predict secondary metabolite gene cluster and anchor genes. In contrast to already available tools, the cluster prediction is based on cluster-specific regulatory patterns. The basic idea is to differentiate cluster from non-cluster genes by regulatory elements within their promoter sequences. They should be enriched in the cluster region in comparison to other parts of the genome.

Inhaltsverzeichnis

Danksagung **5**

Zusammenfassung **7**

Abstract **9**

Abkürzungen **15**

1 Einleitung **17**

1.1 Sekundärmetabolismus und Sekundärmetabolite 17
1.2 Sekundärmetabolit-Gen-Cluster 19
1.3 Ankergene 20
1.4 Ankergenvorhersage 24
1.5 Sekundärmetabolit-Gen-Cluster-Vorhersage 24
1.6 Motivsuche 26
1.6.1 Suche nach bekannten und unbekannten Motiven 27
1.6.2 Konsensus- und profilbasierte Motivsuche 27
1.6.3 Einbeziehung von Daten zusätzlich zur Sequenzinformation . . 30
1.6.4 Auswahl eines geeigneten Motivsuchealgorithmus 31
1.7 Motivbasierte Sekundärmetabolit-Gen-Cluster-Vorhersage 31

2 Manuskripte **33**

2.1 Übersicht zu den Manuskripten 33
2.2 Manuskript 1: »Motif-based method for the genome-wide prediction of eukaryotic gene clusters« 35
2.3 Manuskript 2: »Microevolution of *Candida albicans* in macrophages restores filamentation in a nonfilamentous mutant« 46

2.4 Manuskript 3: »CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes« 65
2.4.1 Ergänzende Materialien zu Manuskript 3 72
2.5 Manuskript 4: »Genome Sequences of Three *Pseudoalteromonas* Strains (P1-8, P1-11, and P1-30), Isolated from the Marine Hydroid *Hydractinia echinata*« 81
2.6 Manuskript 5: »Draft Genome Sequences of Six *Pseudoalteromonas* Strains, P1-7a, P1-9, P1-13-1a, P1-16- 1b, P1-25, and P1-26, Which Induce Larval Settlement and Metamorphosis in *Hydractinia echinata*« 84
2.7 Manuskript 6: »Genetic and metabolic aspects of primary and secondary metabolism of the Zygomycetes« 87

3 Diskussion **113**
3.1 Gründe für die Existenz von Sekundärmetabolit-Gen-Clustern 113
3.2 SMIPS 115
3.3 Anwendung der *de novo* Motivsuche in CASSIS 116
3.3.1 Anzahl und Länge der Promotorsequenzen für die Motivsuche 116
3.3.2 Warum ausschließlich MEME für die Motivsuche genutzt wird 117
3.3.3 Anzahl und Überprüfung der von MEME gefundenen Motive . 118
3.4 CASSIS 120
3.4.1 Unterschiede zwischen CASSIS und MDM 120
3.4.2 Anwendung von CASSIS auf Genome mit Operons 122
3.4.3 Einfluss der Genomsequenz und Genomannotation auf die Sekundärmetabolit-Gen-Cluster-Vorhersage 122
3.5 Nachteile der ähnlichkeitsbasierten Sekundärmetabolit-Gen-Cluster-Vorhersage 124
3.6 Ungeeignete Merkmale für die genaue Vorhersage von Sekundärmetabolit-Gen-Clustern 126
3.6.1 Genexpression 126
3.6.2 DNA-Krümmung und GC-Gehalt 126
3.6.3 Länge der intergenischen Regionen 127
3.7 Zukunft der Sekundärmetabolit-Gen-Cluster-Vorhersage 128

Literaturverzeichnis **131**

Abbildungsverzeichnis **157**

Ehrenwörtliche Erklärung **159**

Abkürzungen

ANR	any number of repetition
antiSMASH	antibiotics & Secondary Metabolite Analysis SHell
bp	Basenpaar(e)
CASSIS	Cluster Assignment by Islands of Sites
ChIP	Chromatin-Immunpräzipitation
DMATS	Dimethylallyltryptophansynthase
FIMO	Find Individual Motif Occurrences
FLI	Fritz-Lipmann-Institut
HKI	Hans-Knöll-Institut
IPR	InterPro Identifier
KEGG	Kyoto Encyclopedia of Genes and Genomes
MEME	Multiple Em for Motif Elicitation
MDM	Motif Density Method
NRPS	nichtribosomale Peptidsynthetase
PKS	Polyketidsynthase
SMGC	Sekundärmetabolit-Gen-Cluster
SMIPS	Secondary Metabolites by InterProScan
SMURF	Secondary Metabolite Unique Regions Finder

TFBS	Transkriptionsfaktorbindestelle
TSS	Transkriptionsstartstelle
ZOOPS	zero or one occurrence per sequence

1 Einleitung

Die Kernthematik der vorliegenden Promotionsarbeit ist die motivbasierte Vorhersage von Sekundärmetabolit-Gen-Clustern (SMGCs). Im Folgenden sollen sowohl der biologische Rahmen dieses Themas, sprich Sekundärmetabolismus (Abschnitt 1.1) und Gen-Cluster (Abschnitt 1.2), als auch die Vorhersage von SMGCs (Abschnitt 1.5) im Allgemeinen, kurz vorgestellt werden. Ein wichtiger Teilaspekt der motivbasierten SMGC-Vorhersage ist die Suche nach Sequenzmotiven. Diese wird in Abschnitt 1.6 vorgestellt.

1.1 Sekundärmetabolismus und Sekundärmetabolite

Unter dem Begriff Metabolismus (Stoffwechsel) werden in der Biologie alle chemischen Prozesse eines Organismus zur Umwandlung von Metaboliten zusammengefasst. Es handelt sich dabei um ein Netzwerk von Stoffwechselwegen, bei denen aus verschiedenen Ausgangsstoffen, über mehrere Zwischenschritte, neue Endprodukte synthetisiert (hergestellt) werden. Die meisten dieser Zwischenschritte werden von Enzymen durchgeführt, andere laufen spontan ab. Eine grafische Darstellung aller bekannten primären und sekundären Stoffwechselwege wird zum Beispiel von KEGG[1] und MetaCyc[2] bereitgestellt.

Albrecht Kossel unterschied 1891 erstmals zwischen Primär- und Sekundärmetabolismus [Bennett und Bentley 1989]. Alle Prozesse, die unmittelbar für das Überleben eines Organismus notwendig sind, werden als Primärmetabolismus bezeichnet. Alle übrigen Prozesse, welche die Überlebenschance unter dem Einfluss wechselnder Umweltbedingungen erhöhen, werden als Sekundärmetabolismus bezeichnet. Sie sind im Allgemeinen nicht notwendig für Wachstum, Entwicklung und Vermehrung [Khaldi

[1] http://www.genome.jp/kegg-bin/show_pathway?map01100, August 2015, Kanehisa und Goto [2000]

[2] http://metacyc.org/META/class-tree?object=Pathways, August 2015, Caspi u. a. [2014]

u. a. 2010]. Der Primärmetabolismus übernimmt allgemeine Aufgaben, die auch in entfernt verwandten Organismen auf identische oder ähnliche Weise ablaufen. Der Sekundärmetabolismus hingegen ist oft auf bestimmte Bedingungen spezialisiert und nur in einer kleinen Gruppe von Lebewesen (wenigen Taxa) konserviert [Keller u. a. 2005]. Eine scharfe Trennung zwischen Primär- und Sekundärmetabolismus ist nicht immer möglich, da die Zuordnung teilweise vom betrachteten Metaboliten und Organismus abhängt [Calvo u. a. 2002].

Sekundärmetabolite (oft auch Naturstoffe genannt) sind die Zwischen- und Endprodukte des Sekundärmetabolismus. Am häufigsten werden sie von Pflanzen, Bakterien und Pilzen synthetisiert. Der Begriff des Sekundärmetaboliten soll anhand dreier Beispiele verdeutlicht werden: Das von *Nicotiana*-Arten (Tabakpflanze) hergestellte Nikotin dient der Abwehr von Insekten [Yamamoto und Casida 1999]. Rapamycin, synthetisiert vom Bakterium *Streptomyces hygroscopicus*, hemmt das Wachstum von Pilzen, mit denen die Bakterien in Nahrungskonkurrenz stehen [VéZina u. a. 1975]. Das von verschiedenen *Penicillium*-Arten (Pinselschimmel) produzierte Penizillin wird zur Verteidigung gegen Bakterien eingesetzt [Fleming 1929]. Diese Beispiele lassen klar erkennen, dass die Produktion der Sekundärmetabolite in einer idealen (»freundlich gesinnten«) Umgebung nicht notwendig ist, in Konkurrenz mit anderen Lebewesen aber maßgeblich zum Überleben beiträgt.

Für den Menschen können Sekundärmetabolite sowohl von positiver als auch von negativer Bedeutung sein. Auf der einen Seite wirkt sich der Kontakt mit giftigen Pflanzen, der Verzehr von ungenießbaren Pilzen oder die Infektion mit einem toxinproduzierenden Bakterium negativ auf den Menschen aus. Auf der anderen Seite wird eine Fülle von Sekundärmetaboliten, oder von ihnen abgeleitete Stoffe, in der Medizin und Pharmazie zur Behandlung von Krankheiten eingesetzt [Newman und Cragg 2012]: Antibiotika gegen Bakterieninfektionen, Antimykotika gegen Pilzinfektionen, Immunsuppressiva zur Unterdrückung des Immunsystems, Statine zur Senkung des Cholesterinspiegels, Proliferationshemmer bei Krebserkrankungen. Durch die Entstehung von multiresistenten Keimen hat vor allem die Erforschung neuer Antibiotika in den letzten Jahren stark an Bedeutung gewonnen [Spellberg u. a. 2008].

Die Grundlagenforschung in diesem Bereich beschäftigt sich unter anderem mit den Fragen, welche Gene bei der Synthese eines bestimmten Sekundärmetaboliten eine Rolle spielen und wie diese Gene reguliert werden. Bei der Beantwortung der Fragen spielen bioinformatische Methoden mittlerweile eine unverzichtbare Rolle.

Die gewonnenen Erkenntnisse können zum Beispiel die Entwicklung eines neuen Antibiotikums ermöglichen oder neue Wege zur Behandlung von Pilzinfektionen aufzeigen.

1.2 Sekundärmetabolit-Gen-Cluster

Gene selbst führen keine Synthese durch, sondern die von ihnen kodierten Proteine und Enzyme. Der Einfachheit halber wird im weiteren Text von Genen geschrieben, auch wenn in manchen Zusammenhängen Proteine gemeint sind.

Die Gene für die Synthese von Sekundärmetaboliten sind oft in Gen-Clustern organisiert [Weber u. a. 2009; Brakhage und Schroeckh 2011]. Als SMGC (Abbildung 1.1, Abbildung 1.2) wird eine Gruppe von Genen bezeichnet, die eng benachbart sind (»kolokalisiert« – sie folgen im Genom aufeinander) und gemeinsam reguliert werden (»koreguliert« – ihre Expression ist gekoppelt). Weiterhin besteht ein Zusammenhang zwischen den von ihnen kodierten Proteinen, weil sie dem gleichen Stoffwechselweg angehören [Keller und Hohn 1997]. Bei Gen-Clustern im weiteren Sinne sind auch weniger strikte Definitionen gebräuchlich [Lemay u. a. 2012]: Manche Autoren fordern zum Beispiel nur den funktionellen Zusammenhang und die räumliche Nähe, nicht aber die gemeinsame Regulierung der Cluster-Gene [Yi u. a. 2007]. In anderen Fällen können Cluster koreguliert und kolokalisiert sein, stehen aber in keinem funktionellem Zusammenhang [Michalak 2008]. Auch die Größe von Gen-Clustern kann, vor allem abhängig von der Definition, sehr verschieden sein: Paare von Genen sind die kleinstmöglichen Gen-Cluster [McGary u. a. 2013]. SMGCs umfassen in der Regel 3 bis 25 Gene (Unterabschnitt 2.4.1, Tabelle S1). Gen-Cluster ohne starke Kolokalisation können sich auf Genomabschnitte von 100.000 Basenpaaren (bp) und mehr erstrecken [Ghanbarian und Hurst 2015]. Die vorliegende Arbeit bezieht sich auf SMGCs, die alle drei Kriterien (*enge* Kolokalisation, Koregulation und funktionellen Zusammenhang) erfüllen.

Gen-Cluster im Allgemeinen finden sich sowohl in prokaryotischen als auch in eukaryotischen Genomen, wobei der Grad der Cluster-Bildung unterschiedlich hoch ist [Blumenthal 1998; Lee und Sonnhammer 2003]. Neben Bakterien, Pflanzen und Pilzen wurden Gen-Cluster auch in Tieren nachgewiesen [Lemay u. a. 2012]. Die vorliegende Arbeit beschäftigt sich hauptsächlich mit eukaryotischen SMGCs von

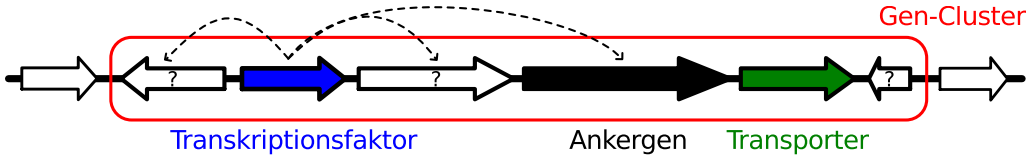

Abbildung 1.1: Schema eines Sekundärmetabolit-Gen-Clusters.

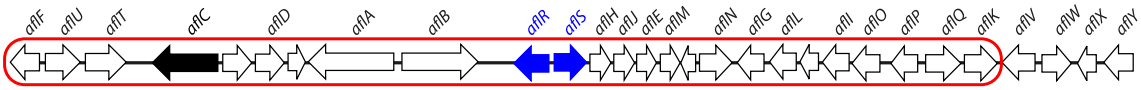

Abbildung 1.2: Beispiel für ein Sekundärmetabolit-Gen-Cluster: Aflatoxin im Genom von *Aspergillus fumigatus*. *aflC*: Polyketidsynthase-Ankergen. *aflR* und *aflS*: cluster-spezifische Transkriptionsfaktoren. Modifiziert nach Amaike und Keller [2011].

(pathogenen) Pilzen.

SMGCs enthalten meist alle Gene, welche für die Synthese des Sekundärmetaboliten notwendig sind. Dies schließt unter anderem Gene für die Modifikationen von Zwischenprodukten, Transporter für den Export aus der Zelle und Transkriptionsfaktoren ein [Brakhage 2013]. Transkriptionsfaktoren sind Proteine, die über eine DNA-Bindedomäne an bestimmte Nukleotidsequenzen binden und die Expression von Genen regulieren [Nguyen und Androulakis 2009]. Die Bindestelle eines Transkriptionsfaktors wird Transkriptionsfaktorbindestelle (TFBS) genannt. Neben cluster-spezifischen Transkriptionsfaktoren, die gezielt ein oder einige wenige Cluster regulieren, gibt es auch nicht-cluster-spezifische (»globale«) Transkriptionsfaktoren [Hoffmeister und Keller 2007]. Im Allgemeinen ist die Zusammensetzung von SMGCs äußerst variabel, da die genannten Arten von Genen Teil des Clusters sein können, aber nicht müssen.

1.3 Ankergene

Eine Konstante unter den Cluster-Genen bilden die Enzyme, welche hauptverantwortlich für die Synthese des Sekundärmetaboliten oder dessen Vorstufe sind. Diese werden als »backbone genes«, »anchor genes«, »signature genes« oder »core genes« bezeichnet, weil sie das Rückgrat der Sekundärmetabolitsynthese darstellen und als Anker oder Mittelpunkt der SMGCs angesehen werden [Khaldi u. a. 2010; Amaike und Keller 2011]. Die beiden am häufigsten bei Pilzen vorkommenden Klassen

von Ankergenen sind Polyketidsynthasen (PKSs) und nichtribosomale Peptidsynthetasen (NRPSs). Kombinationen aus beiden Klassen, sogenannte PKS-NRPS-Hybride, kommen ebenfalls vor [Bergmann u. a. 2007]. Dimethylallyltryptophansynthasen (DMATSs) sind eine weitere Klasse von typischen Ankergenen [Inglis u. a. 2013].

PKSs und NRPSs synthetisieren Polyketide beziehungsweise Peptide durch die schrittweise Verknüpfung von »Grundbausteinen«. Bei PKSs sind dies hauptsächlich Acetyl-Coenzym A und Malonyl-Coenzym A, ähnlich der Fettsäuresynthese [Hopwood 1997]. Bei NRPSs sind es hauptsächlich Aminosäuren, jedoch unabhängig von der typischen Peptidsynthese am Ribosomen [Mootz u. a. 2002].

PKSs bestehen aus verschiedenen Domänen mit unterschiedlichen katalytischen Funktionen (Abbildung 1.3). Alle Domänen einer PKS können entweder von einem oder ein paar wenigen Genen kodiert werden (»Typ I PKS«), oder jede Domäne wird von einem separaten Gen kodiert (»Typ II PKS«). Bei Typ I können außerdem mehrere Domänen zu einem Modul zusammengefasst sein. Die Module werden bei der Synthese des Sekundärmetaboliten nacheinander durchlaufen. Besteht die PKS aus nur einem Modul, so wird dieses mehrmals (iterativ) durchlaufen [Jenke-Kodama u. a. 2005]. NRPSs bestehen ebenfalls aus verschiedenen Domänen, die in Modulen gruppiert sind (Abbildung 1.4). Auch bei NRPSs führen die Module entweder nacheinander (linear) oder in Runden (iterativ) die schrittweise Synthese des Sekundärmetaboliten durch [Mootz u. a. 2002].

Ankergene können anhand ihrer typischen Proteindomänen leicht erkannt werden (Abschnitt 1.4). Der Aufbau und die Wirkungsweise der von ihnen synthetisieren Sekundärmetabolite ist jedoch äußerst vielseitig [Bérdy 2005]. Zum einen können die Grundbausteine verschieden kombiniert werden. Zum anderen befinden sich in SMGCs oft Gene, die den Sekundärmetaboliten weiter modifizieren, unabhängig vom Ankergen. Auf diese Weise entsteht aus den einzelnen Grundbausteinen eine Sekundärmetabolitvorstufe und dann der fertige Wirkstoff.

Die Publikation von Weber [2014] enthält eine umfangreiche Übersicht zu Datenbanken und Programmen, welche Informationen über PKSs, NRPSs und SMGCs enthalten beziehungsweise für deren Vorhersage und Analyse genutzt werden. Zum Beispiel stellen DoBISCUIT [Ichikawa u. a. 2013] und ClusterMine360 [Conway und Boddy 2013] Informationen zu bekannten PKSs, NRPSs und SMGCs zur Verfügung. Beispiele für Programme zur Vorhersage und Analyse von Ankergenen und SMGCs werden in Abschnitt 1.4 beziehungsweise in Abschnitt 1.5 vorgestellt.

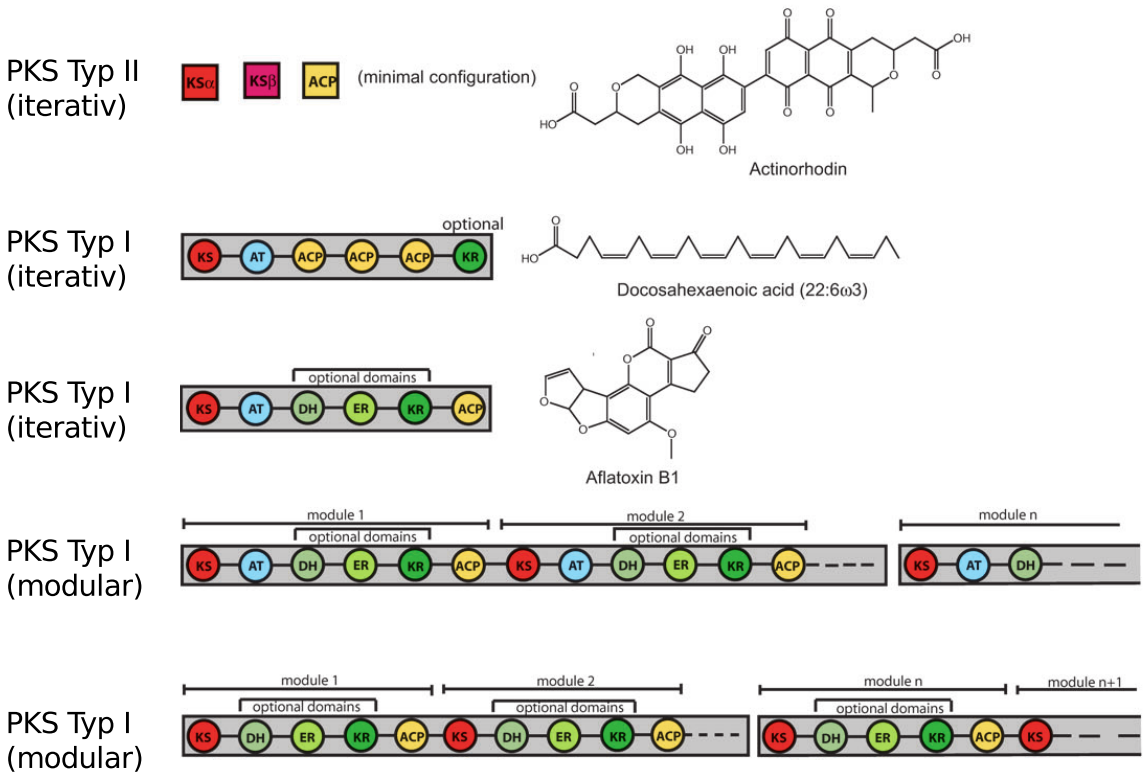

Abbildung 1.3: Schema verschiedener PKS-Typen. Graue Rechtecke: Gene/Proteine. Farbige Kreise: Proteindomänen. Modifiziert nach Jenke-Kodama u. a. [2005].

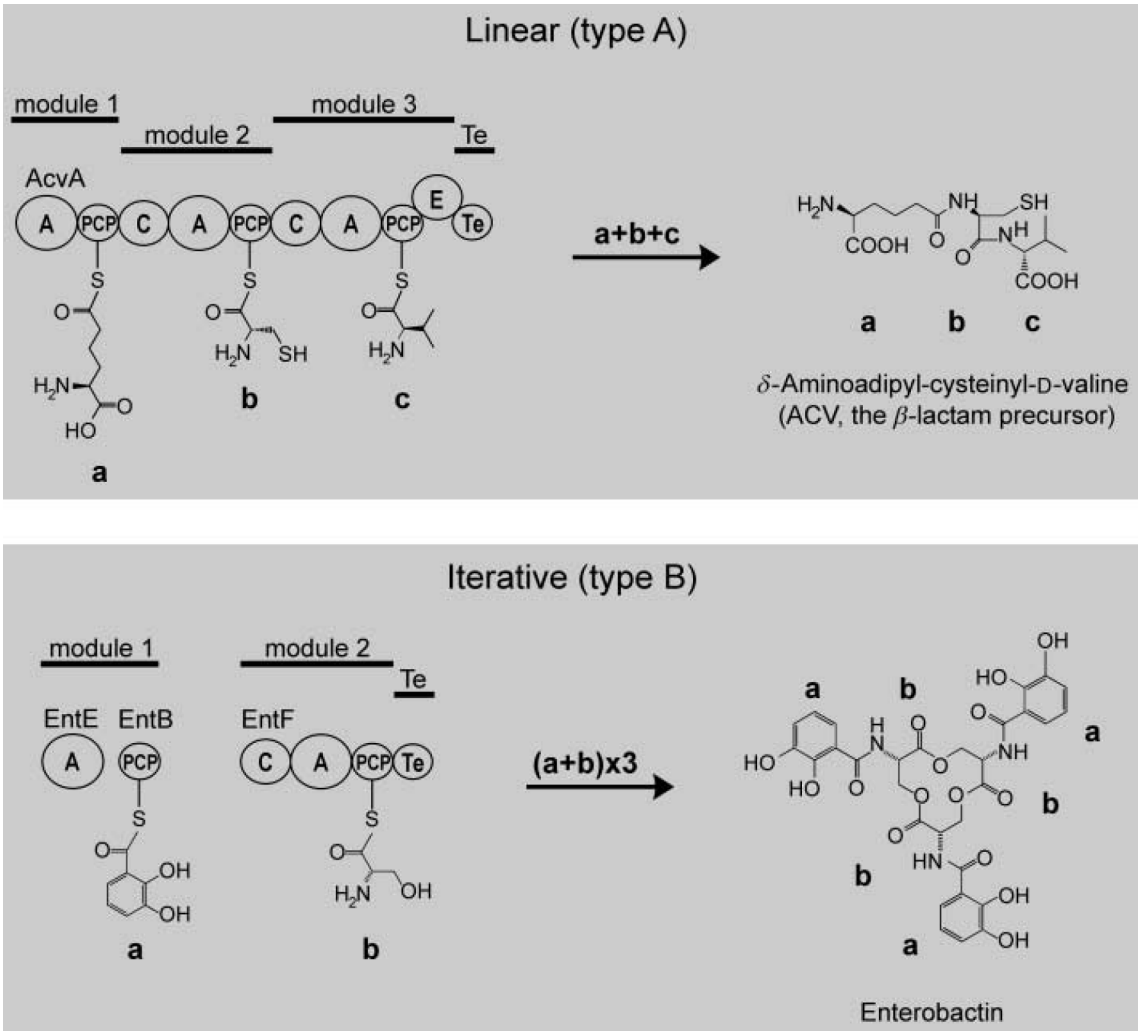

Abbildung 1.4: Schema verschiedener NRPS-Typen. Ellipsen: Proteindomänen. Modifiziert von Mootz u. a. [2002].

1.4 Ankergenvorhersage

Es existieren verschiedene Programme zur Vorhersage und Analyse von Ankergenen. Die meisten Programme erkennen Ankergene anhand ihrer spezifischen Proteindomänen (Abschnitt 1.3). Sie vergleichen die zu analysierende(n) Sequenz(en) mit einer Auswahl an Proteindomänen von bekannten Ankergenen [Weber 2014]. Aufgrund der Aminosäuresequenz sowie der Anordnung und der Anzahl der verschiedenen Domänen können Aussagen über den Typ des Ankergens, die verwendeten Grundbausteine und den synthetisieren Metaboliten getroffen werden.

Auf der einen Seite gibt es Programme, die Ankergene auf bestimmte Eigenschaften hin untersuchen, sie allerdings nicht selbstständig aus einer größeren Menge von Proteinsequenzen herausfiltern können. SEARCHPKS [Yadav u. a. 2003] konnte als eines der ersten Programme die Domänen von modularen PKSs identifizieren und vorhersagen, mit welchen Grundbausteinen das wachsende Polyketid verlängert wird. NRPSpredictor [Rausch u. a. 2005] und NRPSpredictor2 [Röttig u. a. 2011] sind spezialisiert auf die Vorhersage der Substratspezifität von NRPSs (genauer gesagt der Adenylierungsdomänen der NRPSs). Das heißt, sie ermitteln, welche Grundbausteine für die Verlängerung des nichtribosomalen Peptids genutzt werden. Mit der Hilfe einer »Support Vector Machine«[3] können Substratspezifitäten von gänzlich unbekannten Adenylierungsdomänen vorhergesagt werden [Weber 2014].

Auf der anderen Seite identifiziert zum Beispiel MAPSI [Tae u. a. 2009] selbstständig PKS-Gene und analysiert deren Domänen. NP.searcher [Li u. a. 2009] sagt genomweit sowohl PKSs, NRPSs, PKS/NRPS-Hybride als auch deren Syntheseprodukte vorher.

Schließlich ist in machen Programmen zur SMGC-Vorhersage die Möglichkeit zur Ankergenvorhersage integriert. Beispiele dafür finden sich im nächsten Abschnitt.

1.5 Sekundärmetabolit-Gen-Cluster-Vorhersage

Die Vorhersage von SMGCs (Abschnitt 1.2) ist mittlerweile eines der wichtigsten Hilfsmittel bei der Entwicklung neuer pharmazeutischer und medizinischer Wirkstoffe [Weber 2014]. Im Labor können SMGCs durch die systematische Inaktivierung oder

[3] Eine *»Support Vector Machine« (SVM)* [Cortes und Vapnik 1995] ist ein Verfahren zur Mustererkennung, das eine Menge von Objekten in zwei Klassen einteilt, so dass zwischen den Klassen ein möglichst großer Freiraum entsteht.

Überexpression von Genen identifiziert werden [Brakhage und Schroeckh 2011]. Unter Zuhilfenahme bioinformatischer Methoden kann der Anteil zeit- und kostenintensiver Laborarbeit verringert werden. Zusätzlich zur Position der Cluster-Gene, beziehungsweise der Cluster-Grenzen, können teilweise auch komplexere Eigenschaften vorhergesagt werden, zum Beispiel die chemische Struktur des vom Cluster synthetisierten Sekundärmetaboliten. Im Folgenden werden die verschiedenen Methoden zur SMGC-Vorhersage kurz vorgestellt.

Die meisten Methoden zur SMGC-Vorhersage basieren auf einer Sammlung von Proteinen beziehungsweise Proteindomänen, von denen bekannt ist, dass sie häufig in SMGCs vorkommen (ähnlichkeitsbasierte SMGC-Vorhersage, Do und Miyano [2008]; Fedorova u. a. [2012]): Zuerst werden mögliche Ankergene identifiziert (Abschnitt 1.4). Danach werden die Gene vor und nach dem Ankergen untersucht. Kodiert eines der Gene eine Proteindomäne, die auch in der Liste der bekannten SMGC-Proteindomänen vorkommt, wird es als Cluster-Gen markiert. Eine Cluster-Grenze ist erreicht, wenn zum Beispiel eine bestimmte Anzahl an Nicht-Cluster-Genen direkt aufeinander folgt oder die Region zwischen zwei Genen (»intergenische Region«) eine gewisse Länge überschreitet [Khaldi u. a. 2010].

ClustScan [Starcevic u. a. 2008] und CLUSEAN [Weber u. a. 2009]) sind, streng genommen, keine Programme zur SMGC-Vorhersage. Zwar können sie Ankergene finden und weitere Gene in deren Umgebung annotieren, nicht aber die Grenzen unbekannter SMGCs selbstständig ermitteln. Mit SMURF [Khaldi u. a. 2010] hingegen können verschiedene Ankergene in Proteinsequenzen identifiziert und die dazugehörigen Cluster(-grenzen) automatisch vorhergesagt werden. AntiSMASH [Medema u. a. 2011; Blin u. a. 2013; Weber u. a. 2015] ist das bisher umfangreichste Programm für die Vorhersage und Analyse von SMGCs. AntiSMASH kann unter anderem verschiedene Ankergene identifizieren, deren Domänen analysieren, Cluster-Grenzen vorhersagen und nach homologen Clustern suchen.

Zur Modellierung und Suche konservierter Proteindomänen verwenden alle genannten ähnlichkeitsbasierten Methoden Hidden-Markov-Modelle[4]. Diese sind zum Beispiel in HMMER [Eddy 2011] implementiert.

[4] Ein *»hidden Markov model« (HMM)* [Baum und Petrie 1966] ist ein stochastisches Modell eines Systems mit nicht direkt sichtbaren, sondern nur indirekt beobachtbaren Zuständen. Zum Beispiel kann die »wahre« Aminosäuresequenz einer Proteindomäne unbekannt (nicht sichtbar) sein. Nur deren Ausprägung in den verschiedenen Proteinen ist beobachtbar.

Weiterhin können SMGCs mit der Hilfe von Genexpressionsdaten vorhergesagt werden [Andersen u. a. 2013]. Dabei werden alle benachbarten und gleichzeitig exprimierten Gene einem Cluster zugeordnet. Das Computerprogramm MIDDAS-M [Umemura u. a. 2013] setzt diese Idee um.

Takeda u. a. [2014] stellten einen ankergenunabhängigen Algorithmus zur Vorhersage von SMGCs vor. Ihm liegt die Beobachtung zu Grunde, dass sich die Anordnung und Sequenz von Genen verwandter Cluster (Synthese ähnlicher Sekundärmetabolite) tendenziell ähnlicher ist als die der umliegenden Gene.

1.6 Motivsuche

Die Motivsuche ist ein unverzichtbarer Bestandteil der motivbasierten SMGC-Vorhersage. Diese Variante der SMGC-Vorhersage und Alternative zur ähnlichkeitsbasierten SMGC-Vorhersage wird in Abschnitt 1.7 vorgestellt. Im folgenden Abschnitt wird zunächst auf die Motivsuche selbst, unabhängig von der SMGC-Vorhersage, eingegangen.

Ein (Sequenz-)Motiv ist ein wiederkehrendes Muster in einer Nukleotidsequenz[5]. Die genaue Abfolge der Nukleotide des Musters kann von Vorkommen zu Vorkommen leicht verschieden sein. DNA-Bindestellen von Transkriptionsfaktoren werden als (DNA-Binde-)Motive bezeichnet.

Zur Ermittlung von Motiven im Labor müssen teure und zeitaufwendige Methoden, wie zum Beispiel »Electrophoretic Mobility Shift Assays« [Garner und Revzin 1981; Hellman und Fried 2007] oder »DNase I Footprints« [Galas und Schmitz 1978], durchgeführt werden. Für eine exakte Beschreibung der Motive müssen außerdem die Ergebnisse mehrerer Versuche verglichen werden. Die ersten DNA-Bindemotive wurden per Hand ermittelt [Rosenberg und Court 1979]. Da dies mit steigender Anzahl der Sequenzen nicht mehr möglich war, wurden Computerprogramme für die Motivsuche entwickelt. Diese trugen dazu bei, den experimentellen und manuellen Aufwand deutlich zu verringern.

Aus algorithmischer Sicht müssen auch bei der computergestützten Motivsuche einige Hindernisse überwunden werden [Sandve und Drabløs 2006]. Zum Beispiel erschweren die hohe Variabilität und geringe Länge vieler Motive deren Unterscheidung

[5] analog für den gesamten Text: Aminosäuresequenz

von den umgebenden Nukleotidsequenzen [Tompa u. a. 2005].

1.6.1 Suche nach bekannten und unbekannten Motiven

Nach Stormo [2000] kann die Motivsuche in die zwei Teilgebiete »Scannen« und *de novo* Motivsuche unterteilt werden. »Scannen« bedeutet: Das Motiv, zum Beispiel eine TFBS, ist bereits bekannt und es wird nach Vorkommen dieses Motivs in einer Menge von Sequenzen gesucht wird [Bulyk 2004]. Die *de novo* Motivsuche beschäftigt sich mit der Vorhersage von zuvor unbekannten Motiven in einer Menge von Sequenzen. Dieses ist im Allgemeinen das schwierigere der beiden Probleme, da zu Beginn der Suche keine Informationen, wie Länge oder ungefähre Abfolge der Nukleotide des Motivs, zur Verfügung stehen.

Korn u. a. stellten 1977 eines der ersten Computerprogramme zur *de novo* Motivsuche vor. Seitdem ist die Anzahl der verfügbaren Methoden auf über 100 gestiegen. Eine ebenfalls nicht geringe Anzahl an Übersichtsartikeln beschäftigt sich mit dem Vergleich und der Bewertung der verschiedenen Methoden, zum Beispiel Stormo [2000], Pavesi u. a. [2004], Jensen u. a. [2004], Tompa u. a. [2005], Sandve und Drabløs [2006], Sandve u. a. [2007], Klepper u. a. [2008], Nguyen und Androulakis [2009] und Zambelli u. a. [2013]. Die verschiedenen Methoden unterscheiden sich unter anderem durch die unterschiedliche Handhabung folgender Fragen: Wie wird das Motiv dargestellt? Auf welche Weise kann das Motiv gefunden beziehungsweise vom Hintergrund unterschieden werden? Welche Parameter, wie zum Beispiel die Motivlänge, müssen angegeben werden? Außerdem gibt es große Unterschiede bei der Komplexität der Methoden, sprich dem Berechnungsaufwand [Sandve und Drabløs 2006].

1.6.2 Konsensus- und profilbasierte Motivsuche

Wie bereits beschrieben, ist ein Motiv ein wiederkehrendes Muster in einer Menge von Nukleotidsequenzen (Abschnitt 1.6). Die einzelnen Vorkommen des Motivs können sich in ihrer Sequenz leicht unterscheiden. Daher wird eine Darstellungsform benötigt, welche die Sequenz mehrerer Vorkommen zusammenfasst. Anhand der Darstellung (Modellierung) kann die *de novo* Motivsuche in konsensusbasierte und Alignment- oder profilbasierte Methoden unterteilt werden [Stormo 2000; Pavesi u. a. 2004]. Andere gebräuchliche Namen für Profil sind in diesem Zusammenhang

```
TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT
```

`TATAAT`	Konsensussequenz
`TATRNT`	alternative Konsensussequenz mit Platzhaltern

Abbildung 1.5: Beispiel für eine Konsensussequenz. Platzhalter: R = G/A und N = G/A/T/C [Cornish-Bowden 1985]. Modifiziert nach Stormo [2000].

Position / Nukleotid	1	2	3	4	5	6
A	−38	+19	+1	+12	+10	−48
C	−15	−38	−8	−10	−3	−32
G	−13	−48	−6	−7	−10	−48
T	+17	−32	+8	−9	−6	+19

Abbildung 1.6: Beispiel für ein (Sequenz-)Profil mit Gewichten, analog zur Konsensussequenz `TATAAT`. Modifiziert nach Stormo [2000].

»frequency matrix«, »position-specific score matrix« oder »position weight matrix« [Pavesi u. a. 2004]. Werden mehrere Vorkommen eines Motivs in einer Konsensussequenz zusammengefasst, so steht an einer Position der Konsensussequenz das Nukleotid, welches an der gleichen Position in allen Vorkommen am häufigsten auftritt (Abbildung 1.5). Ein (Sequenz-)Profil hingegen ist eine Matrix. Die Zeilen der Matrix entsprechen den möglichen Nukleotiden (A, C, G, T) und die Spalten den Positionen im Motiv (Abbildung 1.6). Im Profil eingetragen sind die Häufigkeiten, Wahrscheinlichkeiten oder Gewichte der verschiedenen Nukleotide an jeder Position, basierend auf einem lokalen Alignment aller Vorkommen des Motivs [Zambelli u. a. 2013]. Eine konsensusbasierte Methode wurde erstmals von Galas u. a. [1985] vorgestellt, eine profilbasierte Methode erstmals von Stormo u. a. [1982].

Die einfachsten konsensusbasierten Methoden suchen nach unveränderlichen Motiven, wie zum Beispiel Verbumculus [Apostolico u. a. 2000]. Andere erlauben eine gewisse Anzahl an Abweichungen, wie Weeder [Pavesi u. a. 2006] und YMF [Sinha und Tompa 2003b].

Konsensussequenzen wurden bald als nicht flexibel genug für die Modellierung biologischer Sequenzen gehalten. Daraufhin wurden verstärkt profilbasierte Methoden entwickelt [Stormo 2000]. Um deren Berechnungsaufwand und Laufzeit zu verringern, wurden Heuristiken[6] angewandt [Pavesi u. a. 2004; Hu u. a. 2005]. Bekannte Beispiele für die profilbasierte Motivsuche sind MEME [Bailey und Elkan 1994] und Gibbs Sampler [Lawrence u. a. 1993; Neuwald u. a. 1995].

MEME ist eine Umsetzung des »Expectation Maximization (EM) Algorithmus« [Lawrence und Reilly 1990] und wichtiger Bestandteil der in Manuskript 1 (Abschnitt 2.2) und Manuskript 3 (Abschnitt 2.4) vorgestellten Programme zur motivbasierten Gen-Cluster-Vorhersage (Abschnitt 1.7). Der EM-Algorithmus besteht aus zwei Schritten: Ausgehend von einem (Start-)Profil wird im E-Schritt bewertet, wie gut verschiedene Abschnitte in den Eingabesequenzen mit dem Profil übereinstimmen. Mit Hilfe dieser Bewertung wird im M-Schritt ein gewichtetes Alignment aller übereinstimmenden Sequenzabschnitte erstellt. Aus diesem Alignment wird ein neues Profil berechnet. Die beiden Schritte werden wiederholt, bis sich die Bewertung des Profils nicht weiter verbessert.

MEME ist seit vielen Jahr ein Teil der »MEME-Suite« [Bailey u. a. 2009, 2015], einer Sammlung von verschiedenen Programmen für die *de novo* Motivsuche, zum »Scannen« und für weitere Analysen rund um Motive und TFBSs. Die Publikation von [Bailey u. a. 2009] ist eine der meistzitierten[7] Texte im Bereich der Motivsuche.

Der ursprüngliche Gibbs Sampler wurde von seinen Autoren mehrfach verbessert (Gibbs Centroid Sampler, Thompson u. a. [2007]), ebenso wurden andere Gibbs-Sampling-Programme entwickelt, wie zum Beispiel AlignACE [Hughes u. a. 2000], Bioprospector [Liu u. a. 2001] und MotifSampler [Thijs u. a. 2002].

Mit den profilbasierten Methoden können keine Abhängigkeiten innerhalb der Motive modelliert werden. Es gibt jedoch Beispiele von Abhängigkeiten zwischen einzelnen Positionen: Badis u. a. [2009]; Eggeling u. a. [2014]. Speziell für die Modellierung von Abhängigkeiten wurde der profilbasierte Ansatz um Markov-Prozesse und Methoden der Bayesschen Statistik ergänzt [Barash u. a. 2003]. Ein konkretes Beispiel hierfür ist VOMBAT [Grau u. a. 2006; Posch u. a. 2007].

[6] Eine *Heuristik* [Pearl 1984] ist ein Verfahren zur Lösung eines Problems, bei dem das Finden einer optimalen Lösung nicht garantiert ist. Ziel ist es, eine »angemessene« Lösung zu finden. Durch Heuristiken können bestimmte Probleme schneller oder überhaupt erst gelöst werden.

[7] laut http://www.webofscience.com, Thomson Reuters, September 2015

TFBSs sind in der Regel nicht zufällig innerhalb der Promotorregionen positioniert [Wray u. a. 2003]. Bayessche Methoden können ebenfalls genutzt werden, um die Präferenz einer TFBS beziehungsweise eines Motivs für eine bestimmte Position zu modellieren. A-GLAM [Kim u. a. 2008] zum Beispiel nutzt sowohl Sequenz- als auch Positionsinformationen für die *de novo* Motivsuche.

1.6.3 Einbeziehung von Daten zusätzlich zur Sequenzinformation

Die Genauigkeit der gefundenen Motive hängt stark von den Eingabesequenzen ab. Um die Auswahl der richtigen Sequenzen nicht allein dem Nutzer zu überlassen, wurden Methoden entwickelt, die selbstständig koregulierte Sequenzen für die Motivsuche aus einer größeren Menge von Sequenzen auswählen. Dies geschieht mit der Hilfe von Genexpressionsdaten. Außerdem wird das Wissen um die verschiedenen Expressionslevel in die Motivsuche einbezogen [Pavesi u. a. 2004]. REDUCE [Roven und Bussemaker 2003], MOTIF REGRESSOR [Conlon u. a. 2003] und RED [Lajoie u. a. 2012] sind Beispiele für diese Art der *de novo* Motivsuche.

Viele Motivsuchealgorithmen sind für die Analyse von Daten einer bestimmten Herkunft, wie »Microarrays«[8] und »ChIPs«[9], optimiert. Zum Beispiel wurden Tawler [Ettwiller u. a. 2007], ChIPMunk [Kulakovskiy u. a. 2010] und Dimont [Grau u. a. 2013] für die Motivsuche in Chip-Sequenzierungsdaten und ähnlichen Techniken entwickelt. Eine besondere Herausforderung für die *de novo* Motivsuche ist hierbei die enorme Anzahl an Eingabesequenzen. Diese erschwert die Entwicklung von effizienten Algorithmen mit moderaten Laufzeiten [Ma u. a. 2012].

Tompa u. a. [2005] und Hu u. a. [2005] schlugen vor, verschiedene Methoden zu nutzen und die Ergebnisse zu vergleichen beziehungsweise zu kombinieren, entweder per Hand oder mit der Hilfe von so genannten »Ensemble-Algorithmen«. CompleteMOTIFs [Kuttippurathu u. a. 2011] zum Beispiel ist eine Plattform, die intern mehrere bekannte Motivsuchealgorithmen, wie MEME und Weeder, nutzt. Mit MotifLab [Klepper und Drabløs 2013] können ebenfalls mehrere Algorithmen, und Daten von verschiedenen experimentellen Techniken, verknüpft werden.

[8] Die Analyse von *Microarrays* [Lockhart und Winzeler 2000] ist eine Methode zur Bestimmung der relativen Genexpression.

[9] Die *Chromatin-Immunpräzipitation (ChIP)* [Collas und Dahl 2008] ist eine Methode zur Bestimmung von Protein-DNA-Interaktionen.

1.6.4 Auswahl eines geeigneten Motivsuchealgorithmus

Die große Anzahl an verschiedenen Methoden erschwert die Auswahl eines geeigneten Motivsuchealgorithmus [Sandve u. a. 2007; Klepper u. a. 2008]. Die eine beste Methode gibt es nicht [Pavesi u. a. 2004; Tompa u. a. 2005]. Bei manchen Datensätzen funktionieren konsensusbasierte Methoden besser, bei anderen die profilbasierten Methoden [Sinha und Tompa 2003a; Sandve u. a. 2007]. Auf keinen Fall sollten perfekte und uneingeschränkt biologisch relevante Ergebnisse von einem Motivsuchealgorithmus erwartet werden [Simcha u. a. 2012]. Im Gegenteil, auf dem Gebiet der computergestützten Motivsuche gibt es nach wie vor großen Spielraum für Verbesserungen, zum Beispiel durch die Einbeziehung von epigenetischen Informationen [Zambelli u. a. 2013]. Für die in dieser Arbeit vorgestellte motivbasierte SMGC-Vorhersagemethode (Abschnitt 1.7) wurde MEME genutzt. Vor- und Nachteile dieser Entscheidung werden in Unterabschnitt 3.3.2 und Unterabschnitt 3.3.3 diskutiert.

1.7 Motivbasierte Sekundärmetabolit-Gen-Cluster-Vorhersage

Die motivbasierte SMGC-Vorhersage ist ein zur ähnlichkeitsbasierten Vorhersage komplementärer Ansatz und die Kernthematik der vorliegenden Arbeit. Dieser Ansatz beruht in erster Linie auf der Annahme der Koregulation, macht sich aber auch die Kolokalisation der Cluster-Gene zunutze (Abschnitt 1.2). Die Grundidee der motivbasierten SMGC-Vorhersage ist die folgende: Wenn alle oder die meisten Gene eines SMGCs koreguliert sind, dann müssen die Promotorsequenzen dieser Gene ein Sequenzmotiv enthalten, welches mit einem cluster-spezifischen Transkriptionsfaktor in Verbindung steht. Vorwissen, zum Beispiel über typische Cluster-Gene oder Genexpressionsprofile, ist nicht notwendig. Nur die Position des Ankergens als Ausgangspunkt für die Cluster-Vorhersage muss bekannt sein. Anschließend werden in der näheren Umgebung des Ankergens Motive vorhergesagt (Abschnitt 1.6) und alle Vorkommen des Motivs im Genom ermittelt. Anhand der Verteilung der Motivfunde – idealerweise alle im Cluster und keine oder nur sehr wenige außerhalb – wird versucht, die Grenzen des Clusters so genau wie möglich zu bestimmen. Alle bisher bekannten ähnlichkeitsbasierten Methoden vernachlässigen die Idee der Koregulation.

Die Manuskripte im folgenden Kapitel beschäftigen sich im Detail mit der Entwicklung und Anwendung zweier Methoden zur Vorhersage von Ankergenen und zur motivbasierten Vorhersage von SMGCs. Schließlich werden in der Diskussion in Kapitel 3 einige Aspekte dieser Einleitung nochmals aufgegriffen und vor allem Details der beiden Vorhersagemethoden kritisch diskutiert, die in den Manuskripten nicht oder nur kurz angesprochen werden.

2 Manuskripte

2.1 Übersicht zu den Manuskripten

Manuskript 1 (Abschnitt 2.2) beschäftigt sich mit der Implementierung der neuartigen Methode MDM zur motivbasierten Vorhersage von SMGCs. In Manuskript 2 (Abschnitt 2.3) wird die Verteilung bestimmter TFBSs im Genom des human-pathogenen Pilzes *Candida albicans* untersucht. Manuskript 3 (Abschnitt 2.4) stellt zum einen das Programm SMIPS zur Vorhersage von Ankergenen vor, zum anderen wird das Programm CASSIS beschrieben, welches die Verbesserung und Erweiterung von MDM ist. In Manuskript 4 (Abschnitt 2.5) und Manuskript 5 (Abschnitt 2.6) wird SMIPS auf das Genom von verschiedenen Stämmen des Bakteriums *Pseudoalteromonas* angewandt, um dessen Potential zur Herstellung von Sekundärmetaboliten zu untersuchen. Schließlich wird in Manuskript 6 (Abschnitt 2.7) die Anwendung von SMIPS auf die Genome verschiedener Zygomyzeten (Jochpilze) beschrieben und die Verteilung der vorhergesagten Ankergene diskutiert.

Tabelle 2.1 listet die jeweiligen Arbeitsanteile aller beteiligten Autoren an den Manuskripten 1 bis 6 auf.

Tabelle 2.1: Arbeitsanteile der Autoren an den Manuskripten

Titel	Author	Anteil
Manuskript 1	**Wolf, Thomas**	50 %
»Motif-based method for the genome-wide prediction of	Shelest, Vladimir	10 %
eukaryotic gene clusters«	Shelest, Ekaterina	40 %
Manuskript 2	Wartenberg, Anja	38 %
	Linde, Jörg	8 %
	Martin, Ronny	2 %
	Schreiner, Maria	2 %
	Horn, Fabian	2 %
	Jacobsen, Ilse D.	4 %
	Jenull, Sabrina	2 %
»Microevolution of *Candida albicans* in macrophages	**Wolf, Thomas**	8 %
restores filamentation in a nonfilamentous mutant«	Kuchler, Karl	2 %
	Guthke, Reinhardt	2 %
	Kurzai, Oliver	2 %
	Forche, Anja	3 %
	d'Enfert, Christophe	7 %
	Brunke, Sascha	10 %
	Hube, Bernhard	8 %
Manuskript 3	**Wolf, Thomas**	50 %
»CASSIS and SMIPS: promoter-based prediction of	Shelest, Vladimir	5 %
secondary metabolite gene clusters in eukaryotic	Nath, Neetika	15 %
genomes«	Shelest, Ekaterina	30 %
Manuskript 4	Klassen, Jonathan	30 %
	Rischer, Maja	25 %
	Wolf, Thomas	20 %
»Genome Sequences of Three *Pseudoalteromonas* Strains	Guo, Huijuan	10 %
(P1-8, P1-11, and P1-30), Isolated from the Marine	Shelest, Ekaterina	5 %
Hydroid *Hydractinia echinata*«	Clardy, John	5 %
	Beemelmanns, Christine	5 %
Manuskript 5	Klassen, Jonathan	30 %
	Wolf, Thomas	25 %
»Draft Genome Sequences of Six *Pseudoalteromonas*	Rischer, Maja	20 %
Strains, P1-7a, P1-9, P1-13-1a, P1-16- 1b, P1-25, and	Guo, Huijuan	10 %
P1-26, Which Induce Larval Settlement and	Shelest, Ekaterina	5 %
Metamorphosis in *Hydractinia echinata*«	Clardy, John	5 %
	Beemelmanns, Christine	5 %
Manuskript 6	Voigt, Kerstin	30 %
	Wolf, Thomas	20 %
	Ochsenreiter, Katrin	10 %
»Genetic and metabolic aspects of primary and	Nagy, Gábor	10 %
secondary metabolism of the Zygomycetes«	Kaerger, Kerstin	10 %
	Shelest, Ekaterina	10 %
	Papp, Tamás	10 %

2.2 Manuskript 1: »Motif-based method for the genome-wide prediction of eukaryotic gene clusters«

Status veröffentlicht, Januar 2013

Literaturangabe Wolf, Thomas ; Shelest, Vladimir ; Shelest, Ekaterina: Motif-Based Method for the Genome-Wide Prediction of Eukaryotic Gene Clusters. In: *New Trends in Image Analysis and Processing – ICIAP 2013*, Springer Berlin Heidelberg, September 2013 (Lecture Notes in Computer Science). `https://link.springer.com/chapter/10.1007/978-3-642-41190-8_42`. – ISBN 978–3–642–41189–2 978–3–642–41190–8, 389–398. – DOI 10.1007/978–3–642–41190–8_42

Übersicht Gene, die nach der Zugehörigkeit zu einem bestimmten Stoffwechselweg gruppiert sind, sogenannte Gen-Cluster, wurden bereits mehrfach in eukaryotischen Genomen nachgewiesen. In dieser Arbeit wird eine neue Methode zur *in silico* Vorhersage von Gen-Clustern vorgestellt, die auf der Hypothese der Kolokalisation und Koregulation der betroffenen Gene beruht. Die Grundidee ist, dass Cluster-Gene von Nicht-Cluster-Genen anhand der Bindestellen für Transkriptionsfaktoren in ihren Promotorsequenzen unterschieden werden können. Dazu wird die Dichte der Vorkommen von Sequenzmotiven – möglichen cluster-spezifischen Bindestellen – ermittelt. Diese sollte im Bereich des Clusters höher sein als im restlichen Genom. Außerdem liefert die Methode nützliche Informationen zu den Regulatoren eines Gen-Clusters.

Beiträge TW, VS und ES konzipierten und entwickelten die Methode. TW implementierte die Methode und wandte sie auf bekannte Gen-Cluster an. TW und ES schrieben das Manuskript. TW war für das Layout und die Finalisierung des Manuskriptes zuständig.

[1] Mit Genehmigung durch *Springer Nature*

Motif-Based Method for the Genome-Wide Prediction of Eukaryotic Gene Clusters

Thomas Wolf, Vladimir Shelest, and Ekaterina Shelest*

Leibniz Institute for Natural Product Research and Infection Biology e. V.
Hans-Knöll-Institute (HKI),
Research group Systems Biology / Bioinformatics,
Beutenbergstrasse 11a, 07745 Jena, Germany
ekaterina.shelest@hki-jena.de

Abstract. Genomic clustering of functionally interrelated genes is not unusual in eukaryotes. In such clusters, co-localized genes are co-regulated and often belong to the same pathway. However, biochemical details are still unknown in many cases, hence computational prediction of clusters' structures is beneficial for understanding their functions. Yet, in silico detection of eukaryotic gene clusters (eGCs) remains a challenging task. We suggest a novel method for eGC detection based on consideration of cluster-specific regulatory patterns. The basic idea is to differentiate cluster from non-cluster genes by regulatory elements within their promoter sequences using the density of cluster-specific motifs' occurrences (which is higher within the cluster region) as an additional distinguishing feature. The effectiveness of the method was demonstrated by successful re-identification of functionally characterized clusters. It is also applicable to the detection of yet unknown eGCs. Additionally, the method provides valuable information about the binding sites for cluster-specific regulators.

Keywords: eukaryotic gene clusters, transcription regulation, secondary metabolites, transcription factor binding sites.

1 Introduction

Genomic clustering (co-localization) of functionally interrelated genes in conjunction with co-regulation, although less present than in prokaryotes, has been found in a great variety of eukaryotic species, from yeast to vertebrates [1,2].

The term "gene cluster" can imply various interpretations. In this work, we consider as clusters the sets of co-localized and co-regulated genes, the products of which are presumably functionally connected (e. g., they can belong to the same biochemical or signaling pathway). Thus, the co-localization and co-regulation are the main characteristics of such eGCs and they form the basis of our approach.

* Corresponding author.

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 389–398, 2013.

Clusters of co-expressed genes have been found in higher eukaryotes, such as drosophila and human [3,4]. It has been shown that genes belonging to the same metabolic pathways are localized significantly closer to each other than it can be expected by chance [2]. This was demonstrated for diverse metabolic pathways from the KEGG database. A relatively well investigated class of eGCs represent the clusters of secondary metabolite genes, which are found in fungi, plants, and protists [5]. Secondary metabolites (SMs) are pharmaceutically important substances (e. g., antibiotics, antimycotics, toxins). The genes responsible for their synthesis, modifications, transport, etc., are often organized in clusters [6]. These clusters are characterized by modest sizes (normally not more than 20 genes) and tight co-localization: the genes are immediately adjacent to each other, although the insertions of non-cluster genes are also possible. The expression of SM clusters is often governed by specific regulators [7] and in many cases the specific transcription factor (TF) is embedded in the cluster [8]. Moreover, non-cluster specific (broad) TFs are also involved in the regulation of SM clusters [9] (Fig. 1).

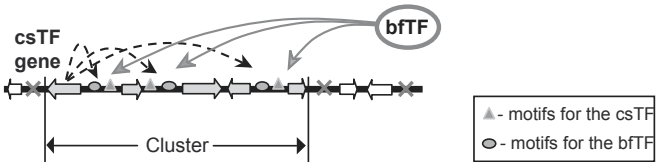

Fig. 1. Regulation of a gene cluster by cluster specific and broad function transcription factors (csTF and bfTF, correspondingly)

There are several ways to predict eGCs genome-wide. One of the first methods was suggested by Lee and Sonnhammer [2] who linked the gene annotations from KEGG with the localization information. The same approach was used in some follow-up works, e. g., in [10], where the authors suggested a method to identify all possible clusters of genes annotated to the same GO term. These methods predict any clusters regardless to their functions and specific features. In the particular cases, like SM clusters, such approaches will give imprecise predictions, mostly because the assignments of genes to pathways are partly or completely unknown.

Another group of cluster-detection methods relies on expression data (microarrays, etc.) [11]. These methods are reliable as long as the data is good, as they provide relatively solid evidence for co-regulation. However, many eGCs, for instance, in fungal genomes are silent under laboratory conditions [6] and it is challenging to experimentally determine the conditions for the cluster induction. Thus, the application of such methods to cryptic clusters is limited.

Some methods have been developed specifically for particular cluster types, e. g., for the SM clusters. Most of the methods developed so far for the detection of SM clusters are similarity based [12,13,14]. Due to the limited number of known clusters that can serve as a template, and also to the possible incorrect assignments of genes to clusters, similarity based methods are error-prone

and tend to overestimate the clusters' lengths. Additionally, these methods do not differentiate closely located (adjacent) eGCs, interpreting them as a single cluster.

These limitations could be circumvented by consideration of sequence characteristics of the cluster regions: GC content and averaged DNA curvative profile [15]. However, not all clusters are characterized by a conserved curvative pattern, which means that a substantial part of them would be skipped by the method if applied to a genome-wide search.

We suggest a novel approach to predict gene clusters based on the density of transcription factor binding site (TFBS) occurrences. In contrast to related tools, our method is not similarity-based. The main idea is that the cluster-specific TFBSs should be enriched in the cluster in comparison to other parts of the genome. Yet, their occurrence outside the cluster is not excluded. We characterize promoters by cluster-specific motif occurrences and consider the density of the motifs as the main feature of the cluster region. The method is applicable to any clusters of co-regulated genes. We demonstrate its usefulness on the example of SM clusters.

2 Results

The presumable co-regulation of the cluster genes presupposes that their promoters share at least one common motif to bind the regulating TF. Ideally, this common motif should be specific to the cluster but not to the surrounding genes (since they are not co-regulated). As the cluster-specific TF (csTF) is assumed not to have ubiquitous functions, its TFBSs should not be widely distributed across the genome. On the other hand, the cluster genes are not necessarily adjacent; "alien" genes inside the cluster may occur (e. g., in [11]). Thus, our requirements for the cluster genes are the following: (i) genes are co-localized; (ii) promoters share at least one common motif; (iii) there can be "gap" genes that do not share the common motif with the rest of the cluster. These requirements allow us to formulate the algorithm to find clusters in a genomic sequence. We call our approach the motif density method (MDM).

2.1 Motif Density Method

The basic idea of the method is that the binding sites for csTFs are enriched in the region of the cluster. Note that we do not exclude their occurrence outside the cluster. Most important, the cluster-specific motifs should be observed in consecutive promoters.

To start the cluster predictions, we need to specify the so-called "anchor" genes. These can be the genes that are already assigned to the pathway in question. In the case of the SM clusters, polyketide synthases (PKSs) or non-ribosomal peptide synthetases (NRPSs) can serve as the anchor genes. PKSs and NRPSs are characterized by a specific set of domains and large size, which makes them relatively easy to detect in genomes.

Step 1: Motif Search. On the first step, all anchor genes are searched and marked in the genomes. Next, an interim set of genes around the anchor gene of interest is selected. Since we do not know how the anchor gene is located relative to the presumable cluster (in the middle or on the edge), we consider several gene sets around the anchor gene not to miss the correct motif: 4/6/8 genes upstream, 4/6/8 genes downstream, and 2 genes up- and downstream the anchor gene. The common motifs are predicted by MEME [16] in the corresponding promoter sequences (−1000/+50 bp around the transcription start site or the whole intergenic region if it is shorter than 1000 bp). Occurrence in the anchor gene promoter is the prerequisite for the further consideration. The best-scoring motif (the one with the lowest score as defined by MEME) out of all considered promoter sets is then searched in all promoter sequences genome-wide.

Step 2: Transforming the Genomic Sequence into the Sequence of Promoters. Counting Occurrences in Frames. On this step, we switch to consideration of promoters as units characterized by the number of occurrences of a particular motif. The order of units follows the order of the corresponding promoters in the genomic sequence. Now instead of the real genomic sequence we consider a string of numbers, which represent the motifs' occurrences in a unit. For instance, if 1 motif was found in the first promoter, 2 motifs in the second, and 0 in the third and fourth promoters, the string will be 1-2-0-0. This number string is scanned by a sliding window (frame) with the step of one unit counting the cumulative number of found motifs per frame. The highest number of occurrences per frame should be obtained for the window coinciding with the cluster. Consideration of different frame lengths allows us to determine the real cluster length.

Step 3: Scoring. To select the optimal frame we apply a scoring system. As the "gap" genes are allowed in the cluster, we allow gaps ("empty" promoters) in the frames but introduce a gap penalty. In this way, we do not forbid the occurrence of small gaps, which are indeed common in clusters, but larger gaps are scored with a penalty that is growing depending on the gap length. The promoters with motifs, on the contrary, add a positive value to the score depending on the number of motifs found.

Let us consider a frame with the length l. In this frame, each promoter i is characterized by the number of found motifs m_i. The consecutive promoters without motifs ($m_i = 0$) form a gap, which is characterized by its length d, the number of gaps in the frame being n.

Then the score S of a frame is calculated as:

$$S = \sum_{i=1}^{l} m_i - \sum_{j=1}^{n} P^{d_j} \quad , \tag{1}$$

where P is the gap penalty and is an adjustable parameter.

The scores are calculated for different frame lengths (normally from 3 to 30, because this is the usual size of the known clusters).

Step 4: Visualization and Selection of the Optimal Frame. The frames are characterized by their score, position, and length. To visualize all characteristics at once, we apply the heat maps (Fig. 2).

2.2 Effectiveness of the Approach

To demonstrate the effectiveness of MDM, we applied it to the re-identification of several functionally characterized SM clusters with known borders. We selected two clusters with characterized regulatory patterns (TFBSs) in order to see if our motif predictions match the real motifs. These chosen examples are the aflatoxin cluster in *Aspergillus flavus* and violaceol cluster in *Aspergillus nidulans*. The latter was also of special interest because it is located in close vicinity to another eGC (orsellinic acid cluster). It was tempting to see if our method is able to separate the two clusters.

The other examples are clusters with characterized products and different patterns of regulation. For instance, the asperfuranone is subject to inter-cluster cross-talk (see Discussion for more details). For all clusters, we compared the predictions of MDM to those of the SMURF tool (Table 1). The gap penalty was set to 1.3 for all examples.

Aflatoxin Cluster in *A. flavus*. Aflatoxin is produced by different Aspergilli [17] and its production is regulated by the csTF AflR, along with several broad function TFs (depending on conditions). The binding sites for AflR have the consensus sequence TCG(N_5)CGA. In *A. flavus*, the cluster spans 21 genes with 15 promoters (AFL2G_07210 to AFL2G_07230). The analysis was run on the genomic sequence from the Broad Institute website [18].

The motif search was performed from scratch in order to confirm the ability of the algorithm to re-identify the real (known) motif. Seven interim sets of promoters around the anchor gene ALF2G_07228 *pksA* (in different arrangements) were submitted to MEME for motif prediction. For each set we could get

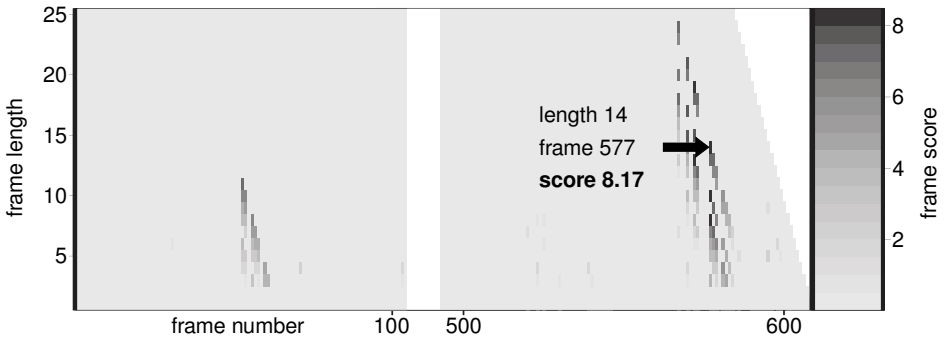

Fig. 2. Heat map for the re-identified aflatoxin cluster (right) and the sub-cluster (left), both on contig 7

common motifs. The motif in the set "8 promoters upstream *pksA*" scored the best and thus was submitted to the genome-wide search. Remarkably, this motif coincided with the AflR TFBS. The AflR motif correctly identified the cluster region with high precision: from AFL2G_07211 to AFL2G_07230 (Fig. 2). The predictions made by the other (non-AflR-like) motifs were much more noisy and failed to detect the cluster.

Violaceol Cluster in *A. nidulans*. The violaceol cluster was described recently [19] and its regulation is yet not well investigated. However, the potential binding sites for the cluster specific regulator were proposed in [19]. MDM was applied to the re-identification of this cluster in the same way as to the aflatoxin cluster, starting with the motif prediction from scratch. The genomic sequence was downloaded from Aspergillus genome database [20]. MDM successfully detected the correct motif (CYCGGAGWWWC) and the correct cluster location (Fig. 3). The length of the cluster is two genes longer than the reported one due to the high number of the csTFBSs in the promoters (Fig. 3). We return to this in the Discussion section. As expected, the orsellinic acid cluster, which is located only five genes apart from the violaceol cluster and which is not regulated by the violaceol csTF, was not detected. In this way, we show the specificity of MDM and its ability to separate closely located clusters.

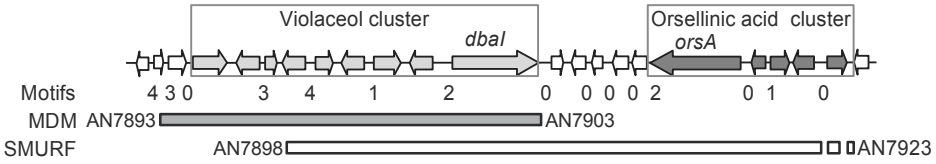

Fig. 3. Re-identification of the violaceol cluster with MDM and SMURF. Coordinates of the real cluster: AN7896 to AN7903.

Asperfuranone Cluster in *A. nidulans*. The regulation of the asperfuranone cluster is a particular case, because the asperfuranone csTF (AfoA) is subject to the regulation by ScpR, the regulator of the NRPS-containing gene cluster *inp*. Under inducing conditions, ScpR triggers AfoA, which in turn induces the expression of the asperfuranone cluster genes (except for AN1031 *afoB*) [21]. Therefore, the *afoA* promoter contains the motif for the ScpR binding [21], whereas the other cluster genes should contain another, not yet described TFBS for AfoA. By the application of MDM we re-identified the cluster nearly perfectly, with expected missing of the *afoA* and *afoB* genes (see also in Discussion).

Aspyridon Cluster in *A. nidulans*, Gliotoxin Cluster in *A. fumigatus*, and WYK-1 cluster in *A. oryzae*. We applied MDM to the re-identification of three more clusters. In all three cases we detected the clusters, although

Table 1. Comparision of SM gene cluster predictions beetween SMURF and MDM

Method	Cluster Start	End	Reference
	Aflatoxin *(Aspergillus flavus)*		
Experimental	AFL2G_07210	AFL2G_07230	[17]
MDM	AFL2G_07211	AFL2G_07230	
SMURF	AFL2G_07219	AFL2G_07248	
	Asperfuranone *(Aspergillus nidulans)*		
Experimental	AN1029	AN1036	[21]
MDM	AN1032	AN1036	
SMURF	AN1029	AN11288[1]	
	Aspyridon *(Aspergillus nidulans)*		
Experimental	AN8408	AN8415	[22]
MDM	AN8401	AN8421	
SMURF	AN8415	AN9243	
	Gliotoxin *(Aspergillus fumigatus)*		
Experimental	AFU6G_09630	AFU6G_09745	[23]
MDM	AFU6G_09630	AFU6G_09785[2]	
SMURF	AFU6G_09580	AFU6G_09740	
	Violaceol *(Aspergillus nidulans)*		
Experimental	AN7896	AN7903	[19]
MDM	AN7893	AN7903	
SMURF	AN7898	AN7923	
	WYK-1 *(Aspergillus oryzae)*		
Experimental	AO090001000009	AO090001000019	[24]
MDM	AO090001000009	AO090001000018	
SMURF	AO090001000009	AO090001000031	

not ideally. The results are presented in Table 1 and discussed in detail in the Discussion section.

3 Discussion

Computational prediction of eukaryotic clusters is especially important when precise information about the corresponding pathways is missing. In such cases, the predicted cluster's structure can point at the involvement of particular enzymes in the pathway and thus be beneficial for the understanding of the pathway's functioning.

Neither of the so far published tools has used the promoter information for the cluster prediction. Since the co-regulation is the basic idea of the cluster

[1] AN11288 is located 2 genes upstream AN1036.
[2] AFU6G_09785 is located 4 genes upstream AFU6G_09745.

definition, we consider the neglect of the promoter information as an oversight. We developed an approach that not only allows to reliably predict the eGCs but also provides information about the potential regulators of the cluster (by description of their TFBSs).

We compared the performance of our method with that of SMURF, the most prominent similarity based approach to SM cluster predictions. SMURF fails to detect the correct borders for most of the clusters and mixes the violaceol cluster with the orsellinic acid cluster reporting them as a single eGC (Fig. 3). MDM gives better or comparable predictions for all examined eGCs and solves the problem of the two adjacent clusters. In the aflatoxin cluster prediction, only one gene of 21 (AFL2G_07210) is missing because the bidirectional promoter between AFL2G_07210 and AFL2G_07209 does not contain the AflR TFBS. This may be reasonable, as AFL2G_07209 does not belong to the cluster and AFL2G_07210 has no assigned cluster function [17]. In the violaceol cluster, two promoters upstream the cluster also shared the specific motif. This does not contradict the experimental data, as the corresponding genes show slight expression under cluster-inducing conditions [19]. In fact, their involvement in the cluster under some specific conditions is not excluded and the function of the csTFBSs deserves additional examination. It remains problematic how to predict clusters with such mosaic regulation. We aim to address this problem in the next versions of MDM.

As mentioned above, the asperfuranone cluster is an interesting case, because its regulator AfoA is induced by a csTF of another cluster. AfoA is shown to induce all cluster genes except for *afoB* [21]. Our findings confirm this experimental result, since the promoter of *afoB* apparently does not contain the AfoA binding motif.

The prediction of the aspyridon cluster by MDM is not perfect, however, it covers the whole cluster, although adding several extra genes up- and downstream of it. Given that SMURF does not find the cluster at all, we consider this result rather good. For the gliotoxin cluster, the left border is found perfectly but on the right side MDM predicts four more genes as cluster members. In such cases (when the promoters have a potential TFBS for a cluster-specific regulator) we cannot exclude a possibility that the cluster is actually longer and those genes can be expressed under some specific conditions. This could be a subject of further experimental investigation. The MDM prediction of the WYK-1 cluster is missing one gene. However, compared to the SMURF result (12 genes more) the prediction of the MDM is closer to the real cluster borders.

The results of the re-identification of the known clusters show that there is space for the improvement of our approach. In many cases, MDM predictions are not perfect. Yet, in the great majority they are better than those made by the similarity-based method, which underscores the higher potential of the motif-based approach.

The genome-wide detection of the csTFBSs can help to discover other genes and even additional clusters regulated by the csTF. Regulatory cross-talk between the clusters has already been described in fungi [21]. In our examples, we

could detect a second peak on the heat map for the AflR motif (Fig. 2). The peak corresponds to a frame in a distant location on the same contig. There is no SM synthase gene in this cluster-like stretch, however, the genes are typical for SM clusters (monooxygenases, methyltransferase, MFS transporters, etc.). There can be two explanations for that: either this is a sub-cluster that is in some way involved in the aflatoxin biosynthetic pathway, or these are the remainings of a damaged cluster that has lost the synthase. In any case, this intriguing sub-cluster deserves further investigation.

To our knowledge, MDM is the first attempt to consider the promoter information in the eGC prediction. We show the high potential of this approach on the examples of the SM clusters, however, the method can be applied to the detection of any eGCs analogous to the SM clusters.

Acknowledgements. This work was financially supported by the Pakt für Wissenschaft und Forschung (2009-2012) and by the International Leibniz Research School for Microbial and Molecular Interactions (ILRS), as part of the excellence graduate school Jena School for Microbial Communication (JSMC), supported by the Deutsche Forschungsgemeinschaft.

References

1. Blumenthal, T.: Gene clusters and polycistronic transcription in eukaryotes. Bioessays 20, 480–487 (1998)
2. Lee, J.M., Sonnhammer, E.L.: Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13, 875–882 (2003)
3. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., Heiqsterkamp, S., van Kampen, A., Versteeg, R.: The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science 291, 1289–1292 (2001)
4. Spellman, P.T., Rubin, G.M.: Evidence for large domains of similarly expressed genes in the Drosophila genome. J. Biol. 1, 5 (2002)
5. Sasso, S., Pohnert, G., Lohr, M., Mittag, M., Hertweck, C.: Microalgae in the postgenomic era: a blooming reservoir for new natural products. FEMS Microbiol. Rev. 36, 761–785 (2012)
6. Brakhage, A.A., Schroeckh, V.: Fungal secondary metabolites - strategies to activate silent gene clusters. Fungal Genet. Biol. 48, 15–22 (2011)
7. Keller, N.P., Hohn, T.M.: Metabolic Pathway Gene Clusters in Filamentous Fungi. Fungal Genet. Biol. 21, 17–29 (1997)
8. Brakhage, A.A.: Regulation of fungal secondary metabolism. Nat. Rev. Microbiol. 11, 21–32 (2013)
9. Hoffmeister, D., Keller, N.P.: Natural products of filamentous fungi: enzymes, genes, and their regulation. Nat. Prod. Rep. 24, 393–416 (2007)
10. Yi, G., Sze, S.H., Thon, M.R.: Identifying clusters of functionally related genes in genomes. Bioinformatics 23, 1053–1060 (2007)
11. Schroeckh, V., Scherlach, K., Nützmann, H.W., Shelest, E., Schmidt-Heck, W., Schuemann, J., Martin, K., Hertweck, C., Brakhage, A.A.: Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in *Aspergillus nidulans*. Proc. Natl. Acad. Sci. USA 106, 14558–14563 (2009)

12. Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., Fedorova, N.D.: SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet. Biol. 47, 736–741 (2010)
13. Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., Breitling, R.: antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39, W339–W346 (2011)
14. Fedorova, N.D., Moktali, V., Medema, M.H.: Bioinformatics approaches and software for detection of secondary metabolic gene clusters. Methods Mol. Biol. 944, 23–45 (2012)
15. Do, J.H., Miyano, S., The, G.C.: window-averaged DNA curvature profile of secondary metabolite gene cluster in *Aspergillus fumigatus* genome. Appl. Microbiol. Biotechnol. 80, 841–847 (2008)
16. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S.: MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208 (2009)
17. Amaike, S., Keller, N.P.: Aspergillus flavus. Annu. Rev. Phytopathol. 49, 107–133 (2011)
18. Aspergillus Comparative Sequencing Project, Broad Institute of Harvard and MIT, `http://www.broadinstitute.org/`
19. Gerke, J., Bayram, O., Feussner, K., Landesfeind, M., Shelest, E., Feussner, I., Braus, G.H.: Breaking the silence: protein stabilization uncovers silenced biosynthetic gene clusters in the fungus *Aspergillus nidulans*. Appl. Environ. Microbiol. 78, 8234–8244 (2012)
20. Arnaud, M.B., Chibucos, M.C., Costanzo, M.C., Crabtree, J., Inglis, D.O., Lotia, A., Orvis, J., Shah, P., Skrzypek, M.S., Binkley, G., Miyasato, S.R., Wortman, J.R., Sherlock, G.: The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community. Nucleic Acids Res. 38, D420–D427 (2010)
21. Bergmann, S., Funk, A.N., Scherlach, K., Schroeckh, V., Shelest, E., Horn, U., Hertweck, C., Brakhage, A.A.: Activation of a silent fungal polyketide biosynthesis pathway through regulatory cross talk with a cryptic nonribosomal peptide synthetase gene cluster. Appl. Environ. Microbiol. 76, 8143–8149 (2010)
22. Bergmann, S., Schümann, J., Scherlach, K., Lange, C., Brakhage, A., Hertweck, C.: Genomics driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. Nat. Chem. Biol. 3, 213–217 (2007)
23. Gardiner, D.M., Howlett, B.: Bioinformatic and expression analysis of the putative gliotoxin biosynthetic gene cluster of *Aspergillus fumigatus*. FEMS Microbiol. Lett. 248, 241–248 (2005)
24. Imamura, K., Tsuyama, Y., Hirata, T., Shiraishi, S., Sakamoto, K., Yamada, O., Akita, O., Shimoi, H.: Identification of a Gene Involved in the Synthesis of a Dipeptidyl Peptidase IV Inhibitor in *Aspergillus oryzae*. Appl. Environ. Microbiol. 78, 6996–7002 (2012)

2.3 Manuskript 2: »Microevolution of *Candida albicans* in macrophages restores filamentation in a nonfilamentous mutant«

Status veröffentlicht, Dezember 2014

Literaturangabe Wartenberg, Anja ; Linde, Jörg ; Martin, Ronny ; Schreiner, Maria ; Horn, Fabian ; Jacobsen, Ilse D. ; Jenull, Sabrina ; Wolf, Thomas ; Kuchler, Karl ; Guthke, Reinhard ; Kurzai, Oliver ; Forche, Anja ; d'Enfert, Christophe; Brunke, Sascha ; Hube, Bernhard: Microevolution of *Candida albicans* in macrophages restores filamentation in a nonfilamentous mutant. *PLOS Genetics* 10 (2014), Dezember, Nr. 12, e1004824. `http://dx.doi.org/10.1371/journal.pgen.1004824`. – DOI 10.1371/journal.pgen.1004824. – ISSN 1553–7404

Übersicht *Candida albicans* und andere humanpathogene Pilze können eine Art »Mikroevolution« durchlaufen. Diese Arbeit zeigt, dass eine nichtfilamentöse Mutante in der Lage ist, innerhalb kürzester Zeit die Fähigkeit zur Bildung von Hyphen zurückzuerlangen. Eine Genom- und Transkriptomanalyse ergab, dass der Austausch eines einzelnen Nukleotides letztendlich das Hyphenwachstum reaktivierte, indem der in der Mutante defekte Signalweg umgangen wurde. Eine genomweite Analyse der Promotorsequenzen ergab, dass Bindestellen des Nrg1-Proteins, einem Unterdrücker des Hyphenwachstums, in hochregulierten Genen zwar häufiger vorkommen als im restlichen Genom, jedoch das Gen NRG1 im evolvierten Stamm weniger stark exprimiert wird, als in der ursprünglichen Mutante vor der Mikroevolution.

Beiträge AW, RM, IDJ, KK, OK, AF, Cd, SB und BH konzipierten und planten die Experimente. AW, RM, MS, IDJ, SJ, AF und Cd führten die Experimente durch. AW, JL, RM, FH, IDJ, SJ, TW, AF, Cd und SB waren für die Datenanalyse zuständig. TW analysierte die Sequenzdaten in Hinblick auf Nrg1-Bindestellenmotive und deren Verteilung im Genom von *Candida albicans*. IDJ, AF, RG, OK, Cd, SB und BH stellten Material und Analysewerkzeuge zur Verfügung. AW, RM, SJ, AF, Cd, SB, RG und BH schrieben die Publikation.

OPEN ACCESS Freely available online

PLOS | GENETICS

Microevolution of *Candida albicans* in Macrophages Restores Filamentation in a Nonfilamentous Mutant

Anja Wartenberg[1], Jörg Linde[2], Ronny Martin[3], Maria Schreiner[1], Fabian Horn[2], Ilse D. Jacobsen[1,4], Sabrina Jenull[5], Thomas Wolf[2], Karl Kuchler[5], Reinhard Guthke[2], Oliver Kurzai[3], Anja Forche[6], Christophe d'Enfert[7,8], Sascha Brunke[1,9☯], Bernhard Hube[1,9,10*☯]

1 Department of Microbial Pathogenicity Mechanisms, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knoell Institute Jena (HKI), Jena, Germany, **2** Research Group Systems Biology & Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knoell Institute Jena (HKI), Jena, Germany, **3** Septomics Research Center, Friedrich Schiller University and Leibniz Institute for Natural Product Research and Infection Biology –Hans Knoell Institute, Jena, Germany, **4** Research Group Microbial Immunology, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knoell Institute Jena (HKI), Jena, Germany, **5** Medical University Vienna, Max F. Perutz Laboratories, Department of Medical Biochemistry, Vienna, Austria, **6** Department of Biology, Bowdoin College, Brunswick, Maine, United States of America, **7** Institut Pasteur, Unité Biologie et Pathogénicité Fongiques, Département Génomes et Génétique, Paris, France, **8** INRA, USC2019, Paris, France, **9** Integrated Research and Treatment Center, Sepsis und Sepsisfolgen, Center for Sepsis Control and Care (CSCC), Universitätsklinikum Jena, Germany, **10** Friedrich Schiller University, Jena, Germany

Abstract

Following antifungal treatment, *Candida albicans*, and other human pathogenic fungi can undergo microevolution, which leads to the emergence of drug resistance. However, the capacity for microevolutionary adaptation of fungi goes beyond the development of resistance against antifungals. Here we used an experimental microevolution approach to show that one of the central pathogenicity mechanisms of *C. albicans*, the yeast-to-hyphae transition, can be subject to experimental evolution. The *C. albicans cph1*Δ/*efg1*Δ mutant is nonfilamentous, as central signaling pathways linking environmental cues to hyphal formation are disrupted. We subjected this mutant to constant selection pressure in the hostile environment of the macrophage phagosome. In a comparatively short time-frame, the mutant evolved the ability to escape macrophages by filamentation. In addition, the evolved mutant exhibited hyper-virulence in a murine infection model and an altered cell wall composition compared to the *cph1*Δ/*efg1*Δ strain. Moreover, the transcriptional regulation of hyphae-associated, and other pathogenicity-related genes became re-responsive to environmental cues in the evolved strain. We went on to identify the causative missense mutation via whole genome- and transcriptome-sequencing: a single nucleotide exchange took place within *SSN3* that encodes a component of the Cdk8 module of the Mediator complex, which links transcription factors with the general transcription machinery. This mutation was responsible for the reconnection of the hyphal growth program with environmental signals in the evolved strain and was sufficient to bypass Efg1/Cph1-dependent filamentation. These data demonstrate that even central transcriptional networks can be remodeled very quickly under appropriate selection pressure.

Citation: Wartenberg A, Linde J, Martin R, Schreiner M, Horn F, et al. (2014) Microevolution of *Candida albicans* in Macrophages Restores Filamentation in a Nonfilamentous Mutant. PLoS Genet 10(12): e1004824. doi:10.1371/journal.pgen.1004824

Editor: Geraldine Butler, University College Dublin, Ireland

Received May 8, 2014; **Accepted** October 15, 2014; **Published** December 4, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Data were deposited at the Gene Expression Omnibus (GSE56174) and can be found in S2 Table.

Funding: This work was supported by the German Federal Ministry of Education and Health (BMBF, www.bmbf.de) Germany, FKZ: 01EO1002 - Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSSC, www.cscc.uniklinikum-jena.de) and the DACH program of the DFG and FWF (DFG HU 528/17-1 & FWF-I-746-B11). AW and TW were also supported by the excellence graduate school Jena School for Microbial Communication (JSMC; www.jsmc.uni-jena.de) and the International Leibniz Research School for Microbial and Biomolecular Interactions (ILRS, www.ilrs.hki-jena.de), respectively. JL was supported by the Deutsche Forschungsgemeinschaft (DFG, www.dfg.de) CRC/Transregio 124 'FungiNet-Pathogenic fungi and their human host: Networks of interaction' (www.funginet.de), subproject INF. IDJ was supported by BMBF 0314108. AF was supported by the National Institute of Allergy and Infectious Diseases, grant R15 AI090633 02 (NIAID, www.niaid.nih.gov). Cd has received funding from the French Government's Investissement d'Avenir program, Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (Grant #ANR-10-LABX-62-IBEID). Work in the KK laboratory was supported by the Austrian Science Foundation FWF by a grant from the Christian Doppler Society. SJ was additionally supported by the FWF (project P-25333-B22). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: bernhard.hube@hki-jena.de

☯ These authors contributed equally to this work.

Introduction

The incidence of invasive fungal infections has steadily increased within the past decades, largely because of a growing population of susceptible individuals, reflecting the progress of modern medicine in prolonging life even with severe underlying diseases and the increasing rate of immuno-deficient patients. One of the most frequently isolated fungi is *Candida albicans*, an ubiquitous and normally harmless commensal of the alimentary tract and mucocutaneous membranes. As an opportunistic pathogen, it can cause superficial infections like oropharyngeal candidiasis, especially in HIV patients, as well as life-threatening

Author Summary

Pathogenic microbes often evolve complex traits to adapt to their respective hosts, and this evolution is ongoing: for example, microorganisms are developing resistance to antimicrobial compounds in the clinical setting. The ability of the common human pathogenic fungus, *Candida albicans*, to switch from yeast to hyphal (filamentous) growth is considered a central virulence attribute. For example, hyphal formation allows *C. albicans* to escape from macrophages following phagocytosis. A well-investigated signaling network integrates different environmental cues to induce and maintain hyphal growth. In fact, deletion of two central transcription factors in this network results in a mutant that is both nonfilamentous and avirulent. We used experimental evolution to study the adaptation capability of this mutant by continuous co-incubation within macrophages. We found that this selection regime led to a relatively rapid re-connection of signaling between environmental cues and the hyphal growth program. Indeed, the evolved mutant regained the ability to filament and its virulence *in vivo*. This bypass of central transcription factors was based on a single nucleotide exchange in a gene encoding a component of the general transcription regulation machinery. Our results show that even a complex regulatory network, such as the transcriptional network which governs hyphal growth, can be remodeled via microevolution.

systemic infections with mortality rates up to 40%, even with current antifungal treatment options [1].

The transition from the commensal to a pathogenic state depends on the microbiota, the host response, and *C. albicans* activities, such as adhesion, secretion of hydrolases, metabolic adaptation, biofilm formation and, importantly, morphological plasticity, which includes the yeast-to-filament transition [2–7]. To survive and thrive in the many different niches inside the host, *C. albicans* must be able to adapt to changing environments and different stresses. In the short term, this occurs primarily by changes in gene expression and translation, and via post-translational modifications, but ultimately microevolutionary processes will play an important role. As a prominent example, White *et al.* [8] have shown that microevolution is the driving force behind the emergence of antifungal drug resistance. They demonstrated the *de novo* appearance of fluconazole resistance in evolving *C. albicans* strains *in vivo* [8]. Furthermore, clinical isolates generally exhibit large genetic variations, and microevolution can be observed both *in vitro* and *in vivo* [9,10], indicating that this process plays an important role in host-pathogen interactions. Therefore, microevolution provides a source of variation for the adaptive response of *C. albicans* to challenging (host) environments.

Different mechanisms account for the generation of new genotypic variants, including point mutations, amplification or deletion of chromosomal segments, chromosomal translocation or inversion, and whole chromosome aneuploidy. These genetic variations can affect expression of single genes or the structure of their encoded proteins as well as whole transcriptional networks via a mechanism known as transcriptional rewiring. In this process, the interaction between promoter regions and their corresponding regulators can be switched to different pairings, which in turn cause new connections to be formed between a signal and a transcriptional response [11,12]. Whereas many studies have explored the underlying mechanisms of drug resistance, the role that microevolution plays in host-pathogen interactions has rarely been investigated: Forche *et al.* [13] found that a *C. albicans* strain, passaged through a mouse host, responded by undergoing chromosome-level genetic variations, which were sufficient to generate new variants of *C. albicans*.

The yeast-to-hyphae transition of *C. albicans* is central for pathogenicity [14,15]. Filamentation plays a pivotal role for adhesion to, invasion into and damage of epithelial and endothelial cells [2,16,17]. Upon internalization by macrophages, *C. albicans* induces host cell death by triggering pyroptosis, a form of programmed cell death [18,19]. However, later in the infection process the yeast-to-hyphae transition contributes to escape from the phagosome [19,20]. Morphology also plays a key role in host recognition [21]. Given the importance of morphology of *C. albicans* for pathogenicity, it is not surprising that the yeast-to-filament transition is induced by a wide range of environmental factors and conditions like high pH, host body temperature, CO_2, starvation and presence of serum, all of which act via several signaling pathways. Among them, the cAMP-dependent protein kinase A (cAMP-PKA) and the mitogen-activated protein kinase (MAPK) pathways, which target the transcription factors Efg1 and Cph1, respectively, play a central role in hyphal formation [22,23]. This is demonstrated by a *cph1*Δ/*efg1*Δ double mutant, which is unable to form hyphae under almost all hyphae-inducing conditions *in vitro* (except agar embedded conditions) and which is probably the most commonly used mutant of *C. albicans* in a wide range of experiments [14,15,22,24].

Due to the central role of the yeast-to-filament transition for *C. albicans* virulence, we used the *cph1*Δ/*efg1*Δ double mutant as a model for evolutionary adaptation. To this end, we performed a series of co-culture passages of this mutant with macrophages. We expected that the hostile environment of the phagosome imposes a high selective pressure on the fungus favoring either intracellular adaptation or return to filamentation in order to escape. We performed phenotypic, transcriptomic and genomic analyses of the pre- and post-passaged strains to elucidate the degree of genetic plasticity of *C. albicans* when facing host stresses. We show that adaptation to macrophages leads to distinct phenotypic differences between the pre- and post-passaged strains with regained filamentation in the latter. As the causative mutation, we identified a heterozygous, non-synonymous single nucleotide exchange in the gene *SSN3*, which encodes the cyclin-dependent kinase of a regulatory module of the Mediator complex. Our results demonstrate that the regulation of the morphological switch in *C. albicans* can be subject to microevolution.

Results

Experimental microevolution causes a reversion of the nonfilamentous phenotype of the *cph1*Δ/*efg1*Δ mutant strain

To determine the ability of *C. albicans* to adapt to stresses inside phagocytes and to test the adaptability of the hyphal regulatory network, we first screened for mutants which are unable to escape from macrophages via filamentation response. We tested multiple *C. albicans* deletion strains with known defects in hyphal formation: strains lacking *RAS1*, *RIM101*, *DFG16*, *TEC1*, *HGC1*, *EED1*, or *UME6* and the avirulent double deletion mutant lacking *CPH1* and *EFG1* [22]. Of these, only the *cph1*Δ/*efg1*Δ double mutant was completely unable to escape from macrophages even after 24 hours, while all other mutants still formed filaments inside the host cell and pierced the phagocyte membrane to some extent (S1A Figure). Microscopy with FITC-labeled *cph1*Δ/*efg1*Δ cells revealed that these cells were viable and

still able to replicate in the yeast form after ingestion by macrophages (S1B Figure).

Therefore, we chose the *cph1*Δ/*efg1*Δ strain for the following microevolution experiment. Cells of the murine macrophage cell line J774A.1 were infected with the *cph1*Δ/*efg1*Δ double mutant at a macrophage-to-fungal ratio of 2:1 and co-incubated. Every 24 hours, non-phagocytosed cells were removed and macrophages were lysed to harvest the phagocytosed cells. A defined fraction of this population was then transferred to a fresh macrophage population.

After 19 passages, a significant morphological alteration became visible, as several phagocytosed cells started to form filaments. These filamenting cells became fixed in the population after additional 23 rounds of co-incubation. This morphologically distinct variant, evolutionary derived from the *cph1*Δ/*efg1*Δ mutant, was termed Evo. The absence of *CPH1* and *EFG1* in the Evo strain was verified by Southern blot analysis (S1C Figure). To exclude temporary or epigenetic effects, the Evo strain was repassaged daily in liquid rich (YPD) medium without any selection pressure by host cells for 14 passages. The phenotype remained stable and no reversal was detected.

To test whether the regained ability to form filaments was restricted to macrophage interactions or observed under additional hypha-inducing conditions, we analyzed the morphology of the Evo strain in the absence of host cells. In the cell culture medium DMEM with 10% serum at 37°C and 5% CO_2, clear filament formation of the Evo strain, but not the *cph1*Δ/*efg1*Δ strain, was observed (Fig. 1). Filamentous growth is associated with highly polarized ergosterol inclusion in membranes, which can be visualized by filipin staining [25]. As shown in Fig. 1, Evo cells grown in the presence of serum exhibited intense filipin staining at the filament tips, equal to the wild type cells. Consistent with the defect in polarized growth, the *cph1*Δ/*efg1*Δ strain showed a more uniform filipin staining. Staining with calcofluor white for morphology analyses showed the expected true hyphae for the wild type and elongated yeasts for the *cph1*Δ/*efg1*Δ strain (Fig. 1). Interestingly, the Evo strain showed heterogeneous cell morphologies, i.e. a mixture of pseudohyphae with constrictions at the septa and true hyphae with parallel-sided walls (Fig. 1). The percentage of the different morphological forms was quantified using the morphological index (MI) [26] of individual cells after 4 and 12 hours of growth in serum (S2A Figure). The MI for *cph1*Δ/*efg1*Δ was <2.5 at both time points, indicating yeast morphology. In contrast, most cells of the Evo strain grew as pseudohyphae (MI 2.5–3.4) after 4 hours, while after 12 hours true hyphae were evident (MI>3.4) in approx. 50% of the population. Both morphologies will be referred to as filaments.

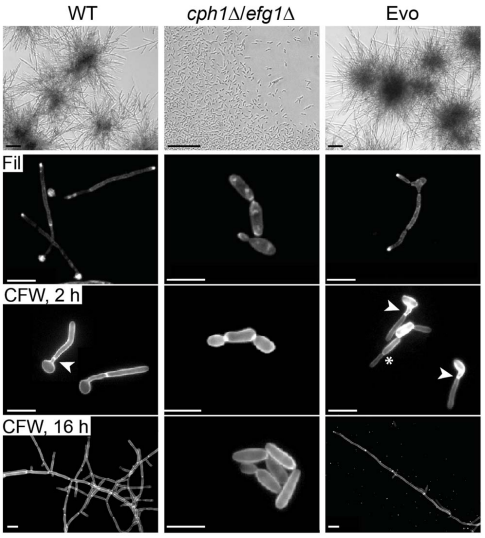

Fig. 1. Co-incubation with macrophages led to regained filamentation in the *cph1*Δ/*efg1*Δ strain. Morphology of wild type (WT), *cph1*Δ/*efg1*Δ and Evo strain after 18 h of incubation in DMEM+ 10% FBS at 37°C and 5% CO_2 demonstrate the re-appearance of filamentation in the Evo strain (scale bar: 100 μm, upper panel). All strains were grown for 4 h for filipin (Fil) staining, and for 2 h or 16 h for calcofluor white (CFW) staining on cover slips, and analyzed by fluorescence microscopy (scale bar: 10 μm, lower panels). Arrow heads highlight septa (true hyphae), while asterisks indicate constrictions (pseudohyphae).
doi:10.1371/journal.pgen.1004824.g001

We then tested different classical hyphae-induction media for *C. albicans* to assess the extent of phenotype reversal to wild type morphology. In response to serum-containing YPD medium with 5% CO_2, the Evo strain initially formed filaments but switched back to yeast growth much earlier than the wild type (S2B Figure). Filamentation (mainly pseudohyphae) also occurred in response to the amino sugar N-acetyl-D-glucosamine as sole carbon source and 5% CO_2 (S2B Figure). Finally, cells of the Evo strain were incubated in serum-containing water at 37°C in atmospheric air. Again, stable filamentation was induced, demonstrating that high CO_2 is not absolutely necessary for filamentation of the Evo strain (S2B Figure). In embedded media at 23°C (S2C Figure.), deletion of *EFG1* causes a hyperfilamentous phenotype [24]. Accordingly, the *cph1*Δ/*efg1*Δ strain was hyperfilamentous under these conditions. Interestingly, while cells of the Evo strain displayed an even more pronounced hyperfilamentous phenotype, it did not undergo filamentation on solid medium at 37°C, as seen in the *cph1*Δ/*efg1*Δ strain (S2D Figure).

In conclusion, our microevolution experiment led to the regained ability of filamentous growth in the *cph1*Δ/*efg1*Δ mutant in response to a diverse range of hyphae-inducing conditions, indicating that microevolutionary events had enabled this strain to bypass the dependency on Cph1 and Efg1 for filamentation in these media.

The Evo strain regained virulence potential

Filamentous growth is an important contributing factor for the escape from macrophages. We therefore determined the amount of Evo cells that escaped from macrophages by piercing through their membranes after 4 h, 6 h and 8 h of co-incubation (Fig. 2A). Both Evo and wild type, but not the *cph1*Δ/*efg1*Δ double mutant, were able to escape from macrophages. However, the piercing rate of the Evo strain was significantly lower than for the wild type at all time points. After 8 h of co-incubation nearly all wild type cells had escaped from the macrophages, but only about 25% of Evo cells. The delay in filamentation and the presence of pseudohyphae in the Evo strain may explain these differences. Next, we assessed the fungus' ability to invade oral epithelial cells. Invasion requires previous adhesion, and the *cph1*Δ/*efg1*Δ strain was almost entirely unable to adhere to epithelial cells (Fig. 2B). Adhesion of the Evo strain was still reduced compared to the wild type, but significantly higher than for the double mutant (Fig. 2B). This is reflected by the invasion capacity of the Evo strain, which was significantly lower than the wild type strain, but substantially

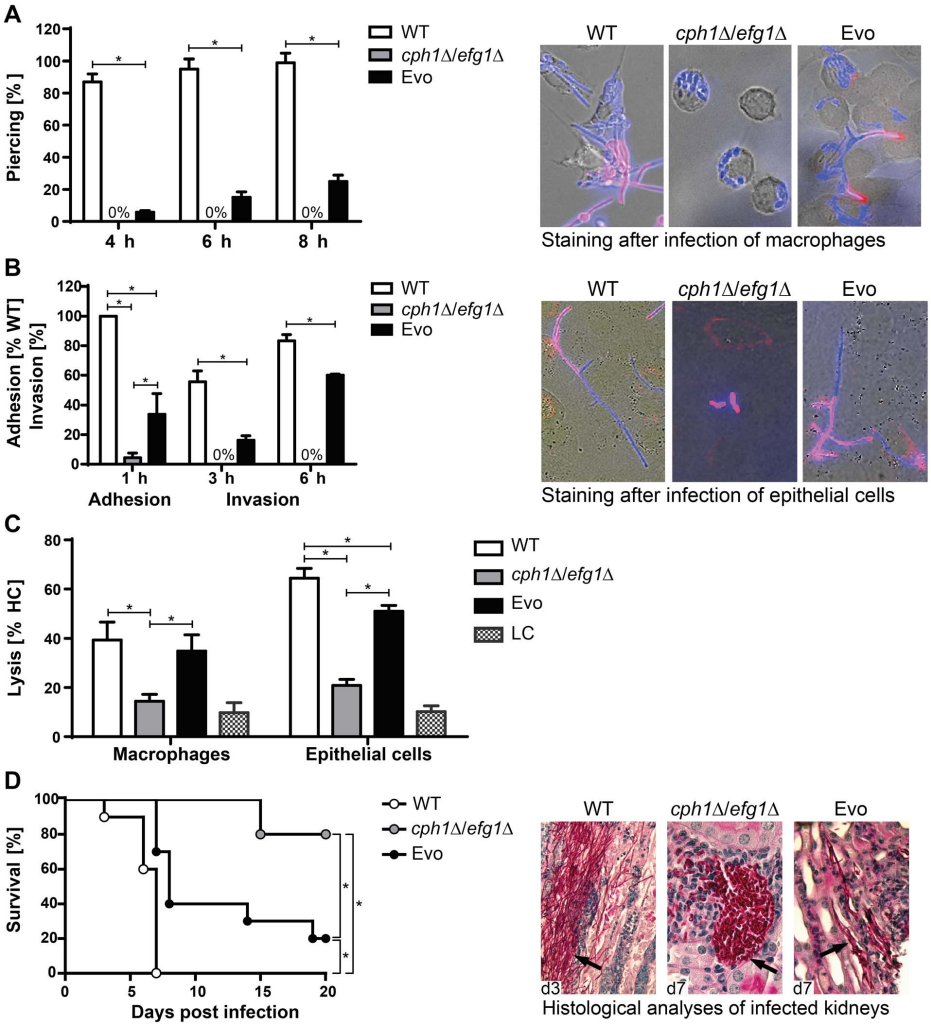

Fig. 2. Characterization of Evo strain interaction with host cells and virulence potential. (**A**) Escape of *C. albicans* cells by piercing of macrophages (J774A.1) after different timepoints (left). Micrographs of strains after 6 h of co-incubation with J774A.1 cells (right). Intracellular *C. albicans* appears blue (CFW), extracellular section of the cells red (Concanavalin A, ConA). Cells of the *cph1Δ/efg1Δ* strain cannot escape from macrophages, while Evo cells regained this property during the evolution experiment. (**B**) Adhesion to and invasion of oral epithelial cells (TR-146). Adhesion values are given as percentage of adherent WT cells (left). Micrographs show filamentation of *C. albicans* WT and Evo strains after 6 h of incubation with TR-146 cells (right). The regained ability to filament enabled the Evo strain to invade epithelial cells. Staining was performed as described in (A). (**C**) Damage to macrophages and epithelial monolayers, determined by lactate dehydrogenase (LDH) assay after 32 h of co-incubation (LC = low control, medium only). WT and Evo strain, but not the *cph1Δ/efg1Δ* strain caused clear damage to both cell types. For piercing, adhesion, invasion and cell damage assay results are given as mean+SD of three independent experiments (*$p<0.05$). (**D**) Survival of BALB/c mice challenged intravenously (left; n = 10/strain). Nearly all mice infected with the Evo strain succumbed to the infection, while almost all animals infected with *cph1Δ/efg1Δ* strain survived (*$p<0.05$). PAS-hematoxylin-stained kidney sections from different days (d) post challenge (right) show fungal cells (arrows) either in the filamentous form (WT and Evo strain) or yeast form (*cph1Δ/efg1Δ* strain).
doi:10.1371/journal.pgen.1004824.g002

higher than the *cph1*Δ/*efg1*Δ strain. Finally, we also investigated the potential of the Evo strain to damage macrophages and epithelial cells by measuring the release of lactate dehydrogenase (LDH). After 32 hours of co-incubation, the Evo strain had damaged macrophages to the same extent as the wild type strain, and epithelial cells to a significantly higher degree than the *cph1*Δ/*efg1*Δ strain (Fig. 2C).

The Evo strain had thus regained abilities putatively relevant for systemic infections. Hence, the virulence of the Evo strain was tested in a murine model of hematogenously disseminated candidiasis. Survival was monitored over a period of 21 days. As predicted, mice infected with the Evo strain showed an intermediate and significantly different survival rate compared to mice infected with the wild type and *cph1*Δ/*efg1*Δ strains (Fig. 2D). Histological examination of kidneys from infected animals revealed that the Evo strain retained its filamentous morphology *in vivo*, even though filaments formed by the Evo strain were shorter than by the wild type, and invasion into deeper layers of the kidney tissue was less pronounced (Fig. 2D).

In summary, the evolved changes in response to macrophages enabled the Evo strain not only to form filaments *in vitro*, but also in contact with host cells, which correlated with a higher virulence potential both *in vitro* and *in vivo*.

The Evo strain expresses hyphal-associated genes and responds to farnesol

Hyphal-associated virulence of *C. albicans* is not only due to filamentation *per se*, but also to the expression of hyphae-associated genes. In order to monitor the expression of typical hyphae-associated genes in the Evo strain, we measured the mRNA levels of *HWP1*, *ECE1* and *ALS3*, all encoding hyphal cell surface proteins, and of *EED1*, a gene that is associated with hyphal cell elongation [27]. An upregulation of all four genes in the Evo strain was confirmed by qRT-PCR after 1 hour of incubation in DMEM+10% FBS at 37°C and 5% CO_2 (Fig. 3A). *HWP1* expression was similar in the Evo and WT strain, whereas *ECE1* and *ALS3* were higher expressed in the WT strain, and *EED1* was more strongly upregulated in the Evo strain. Furthermore, we observed Als3 exposure on the surface of wild type and Evo cells by immunofluorescence, but not on the *cph1*Δ/*efg1*Δ strain (Fig. 3B). This regained cell-surface exposure of the Als3 adhesin [28] is in accordance with the increased adhesion potential of the Evo strain.

We were next interested if the filamentation program can be blocked by the quorum-sensing molecule farnesol. Very low concentrations (1 μM) of farnesol in the medium resulted in a complete repression of filament formation in the Evo strain, whereas wild type cells still formed hyphae (Fig. 3C). Consistently, farnesol treatment led to a dramatic repression of filament-associated gene expression (Fig. 3D). By addition of exogenous dibutyryl-cyclic AMP (db-cAMP) to the farnesol-containing medium, filamentation was rescued in the Evo strain (Fig. 3C). These data suggest a critical role for cAMP signaling in the filamentation process of the Evo strain.

The Evo strain shows wild type levels of filamentation-associated transcription factor gene expression

The yeast-to-filament regulatory network comprises many different transcription factors (TFs). The filament-associated biofilm formation is controlled by a network formed by Bcr1, Tec1, Brg1, Rob1, Ndt80 and Efg1 [29]. Efg1 positively regulates all other TF genes in this network except *ROB1*. We measured the transcription of these central TF genes at 30 min and 60 min after filament induction. As shown in Fig. 4A, we found an at least 1.5-fold upregulation of *ROB1* and *TEC1* after 30 min, and of *BCR1* and *BRG1* at both timepoints in the Evo strain. The wild type, however, showed only an increased expression of *TEC1* at both timepoints and of *BRG1* after 30 min. In contrast, most of these TF genes were down- or scarcely upregulated in the *cph1*Δ/*efg1*Δ strain (Fig. 4A).

Formation of wild type filaments is also regulated in part by *CZF1* under certain conditions [30]. An increased expression of *CZF1*, however, is not the cause for filamentation in the Evo strain. The *CZF1* mRNA levels under serum induction did not greatly differ from the mRNA level in the *cph1*Δ/*efg1*Δ strain (Fig. 4B). In addition, the mRNA level of *UME6*, a key TF gene necessary for the maintenance of filamentation [31], was upregulated in wild type cells at both time points but not in the *cph1*Δ/*efg1*Δ strain (Fig. 4B). Interestingly, *UME6* expression was more than 4-fold upregulated in the Evo strain after 60 min growth in serum-containing medium.

C. albicans possesses an *EFG1* homolog, *EFH1*, and overexpression of this gene is known to induce pseudohyphal growth. In addition, like *EFG1*, *EFH1* is involved in the regulation of expression of filament-associated genes [24]. We found that *EFH1* showed the strongest upregulation (7.4-fold) among the tested TF genes in the Evo strain. However, deletion of *EFH1* did not abolish filamentation of an Evo strain derivative (Fig. 4C). Hence, the filamentation phenotype of the Evo strain was not linked to this TF.

In summary, the Evo strain has regained most of the transcriptional hallmarks of filament production, including the upregulation of the central transcription factor genes *TEC1*, *BRG1* and *UME6*. The few discrepancies to the wild type may partially explain the remaining differences in morphology. However, the late-phase upregulation of *UME6* indicates that the filament maintenance of the Evo strain is similar to the wild type at the transcriptional level. Furthermore, the function of Efg1 was not replaced by Efh1 in the Evo strain.

The cell wall defects of the *cph1*Δ/*efg1*Δ mutant are reverted in the Evo strain

Our data indicated that the Evo strain regained the potential to produce hyphae, showed upregulation of transcription factor genes involved in filamentous growth and other hyphal associated genes, and regained a high virulence potential. The reduced virulence of the *cph1*Δ/*efg1*Δ strain is likely predominantly caused by the filamentation defects, however, Efg1 has also an important role in cell wall architecture [32] and the cell wall is essential for adhesion and invasive growth and thus for pathogenicity [33]. We therefore tested the Evo strain for cell wall defects by treatment with cell wall perturbing agents, i.e. congo red (CR), calcofluor white (CFW) and sodium dodecyl sulfate (SDS). As shown in Fig. 5A, the *cph1*Δ/*efg1*Δ strain was hypersensitive to all tested agents. In contrast, the Evo strain was as resistant as the wild type to CR and CFW, agents that disturb glucan and chitin architecture, respectively. The same phenotypic reversal was observed for the cell membrane disturbing agent SDS, suggesting a loose structure of the cell wall only in the *cph1*Δ/*efg1*Δ strain.

These results indicate that the altered cell wall composition of the *cph1*Δ/*efg1*Δ strain was at least partially restored in the Evo strain. We therefore stained exposed mannan and β-1,3-glucan with fluorescently labeled concanavalin A (ConA) and anti-β-1,3-glucan antibody, respectively (Fig. 5B). Quantification by FACS analysis displayed significantly reduced mannan and increased β-1,3-glucan signals on the surface of the *cph1*Δ/*efg1*Δ strain

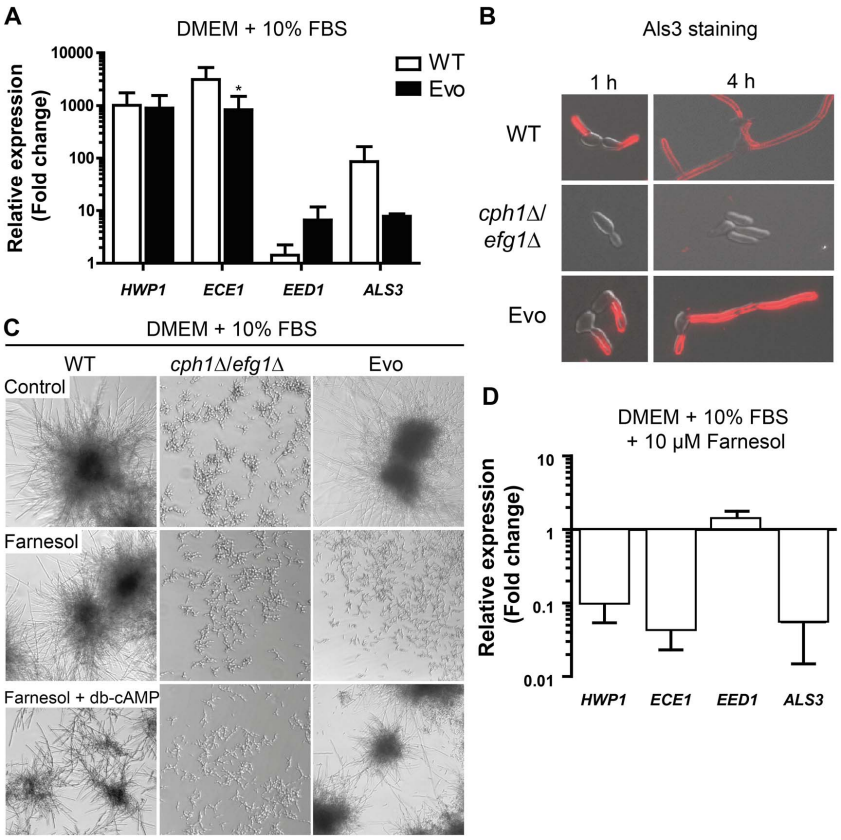

Fig. 3. Analysis of hyphae-associated gene expression, Als3 surface expression and response to farnesol. (**A**) The Evo strain expresses hyphae-associated genes after growth for 1 h at 37°C at 5% CO_2 on a plastic surface similar to WT. Relative gene expression of filament-inducing conditions was compared to yeast promoting conditions (YPD, 30°C) for three independent experiments. Expression was normalized against three housekeeping genes (*ACT1*, *EFB1* and *PMA1*) and data are shown as mean+SD of three biological experiments (*$p<0.05$). (**B**) Immunofluorescence micrographs of cells immuno-stained for Als3 after growth in DMEM+10% FBS at 37°C and 5% CO_2 on cover slips. Wild type (WT) and Evo cells are Als3-positive, while *cph1Δ/efg1Δ* cells show no signal (representative samples). (**C**) Morphogenetic response to farnesol treatment alone or in combination with exogenous dibutyryl-cyclic AMP (db-cAMP). All strains were exposed to either methanol (control), 1 μM farnesol or 1 μM farnesol+ 10 mM db-cAMP and incubated at 37°C and 5% CO_2 for 18 h (representative pictures from three independent experiments are shown). Note that in Evo cells inhibition of filamentation by farnesol treatment was completely abrogated when db-cAMP was added. (**D**) Repression of hyphae-associated gene expression in the Evo strain by 10 μM farnesol. Expression was normalized against three housekeeping genes (*ACT1*, *EFB1* and *PMA1*). The fold change in expression relative to filament-inducing conditions alone is shown as mean+SD of three biological experiments.
doi:10.1371/journal.pgen.1004824.g003

compared to the wild type strain. The Evo strain showed an intermediate mannan and wild type-like glucan exposure.

The two MAP kinases, Cek1 and Mkc1, become activated in wild type *C. albicans* upon treatment with cell wall disturbing agents [34–36]. After treatment with CR, both Cek1 and Mkc1 were phosphorylated in the Evo strain but not in the *cph1Δ/efg1Δ* strain (Fig. 5C). Unusually, only phosphorylated Mkc1 could be detected in the wild type strain, which may be due to changes in the CR treatment protocol compared to previous experiments performed by another group [34]. However, these results show that the Evo strain regained the ability to phosphorylate Mkc1 and Cek1 in response to cell wall stress.

Global transcriptional analysis of the evolved strain by RNA-Seq

To gain insight into the regulatory program of filamentation in the absence of *CPH1* and *EFG1*, we performed gene expression analysis by RNA sequencing under hyphae- and non-hyphae-inducing conditions. Both, the *cph1Δ/efg1Δ* and the Evo strain, were analyzed with a sequencing depth sufficient to cover the genome 75–350×. Expression (RPKM≥1) was detected for 5,854 of the *C. albicans* open reading frames (94%), as well as for 561 nTARs (novel transcriptionally active regions, [37]), 67 small nuclear RNAs and for 24 tRNAs (see Materials and Methods for details and S2 Table for a complete list of all detected transcripts).

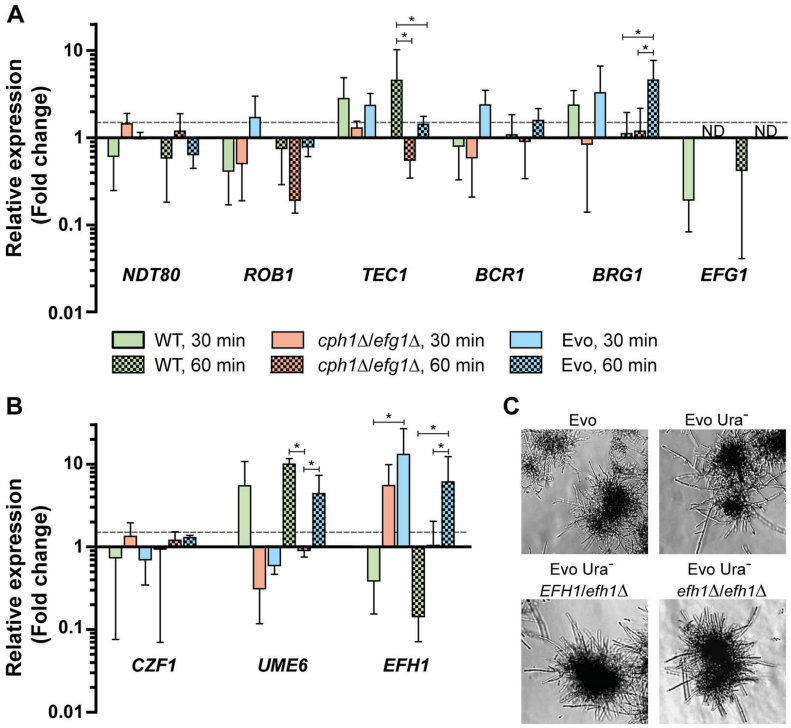

Fig. 4. Expression of transcription factors under filament-inducing conditions. (**A+B**) Relative expression of nine central transcription factor genes in the analyzed strains after growth in DMEM+10% FBS at 37°C and 5% CO_2 on a plastic surface. Fold change between filament-inducing and yeast promoting conditions (YPD, 30°C) is shown, normalized to three housekeeping genes (*ACT1*, *EFB1* and *PMA1*). Means+SD of n = 3 (dotted line indicates threshold at 1.5; *p<0.05). (**C**) Deletion of *EFH1* in the Evo strain did not affect hyphal growth. Cells were incubated for 18 h at 37°C and 5% CO_2 in DMEM+10% FBS (representative pictures).
doi:10.1371/journal.pgen.1004824.g004

Differential expression of selected genes was subsequently validated by qRT-PCR using biological replicates (S3A Figure).

After the transfer to filament-inducing conditions, 379 transcripts were significantly upregulated (≥2-fold, p<0.01) and 279 downregulated in the Evo strain. In the *cph1Δ/efg1Δ* strain, 255 transcripts were up- and 252 downregulated under the same condition. Within the group of upregulated transcripts, 209 genes were induced in both strains, while 46 transcripts were specifically induced in the *cph1Δ/efg1Δ* strain and 170 transcripts specifically in the Evo strain. 186 of the downregulated transcripts were repressed in both strains, whereas 66 and 93 transcripts were specifically repressed in the *cph1Δ/efg1Δ* and Evo strains, respectively (S3B Figure).

We investigated the expression of individual marker genes for filamentation [38] more closely (S3C Figure). As expected, all eight genes of the core filamentation response (*ALS3*, *ECE1*, *DCK1*, *HGT2*, *HWP1*, *IHD1*, *RBT1* and orf19.2457) were significantly upregulated in the Evo strain under filament-inducing conditions. Four of these genes (*ECE1*, *HWP1*, *IHD1* and *RBT1*) and further filament-associated genes, like *ALS1*, *BRG1* and *HGC1* were also upregulated in the non-filamenting *cph1Δ/efg1Δ* strain. Expression of filament-associated genes independent of any morphological transition has previously been described in the *cph1Δ/efg1Δ* mutant [38–40]. However, these genes were expressed at a significantly higher level in the Evo strain compared to the *cph1Δ/efg1Δ* strain under filament-inducing condition (S2 DS5 Table).

Overall, genes most highly expressed (≥5-fold) in the Evo strain under filament-inducing condition are mainly hyphal-associated genes (*HWP1*, *ECE1*, *ALS3*, *RBT1*, *FRG2*, *ALS1* and *IHD1*). Furthermore, the expression of *YWP1*, encoding a yeast-form cell wall protein, is downregulated in the Evo strain, while its expression did not change in the *cph1Δ/efg1Δ* strain. These results suggest that genes associated with *C. albicans* hyphae formation are also associated with filamentation of the Evo strain.

Upregulation (>1.5-fold, p<0.01; S2 DS1 and DS4 Table) of *DCK1*, *LMO1* and *CEK1*, which are required for filamentation under embedded conditions and for cell wall integrity [41], was found solely in the Evo strain. This provides a possible explanation for the hyper-filamentous phenotype under embedded conditions as well as the increased resistance to cell wall perturbants compared to the double mutant (Figs. 1+5).

To determine whether changes in the regulation of effector genes are reflected by an upregulation of specific TF genes, we also

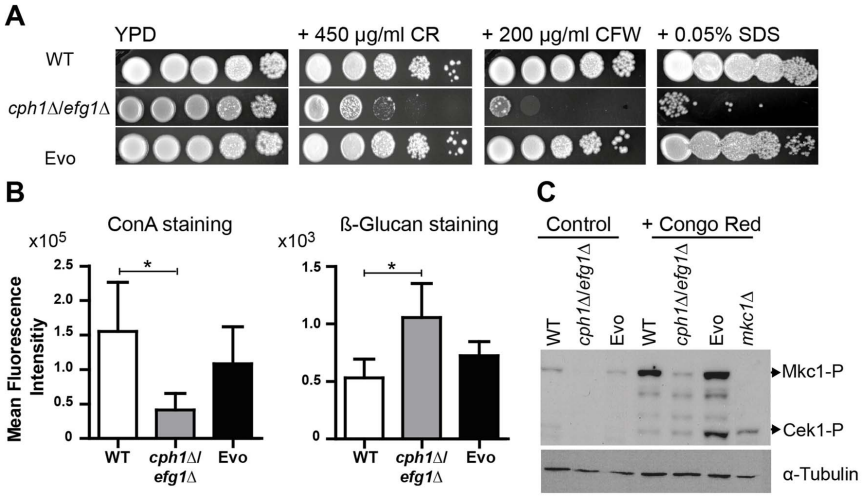

Fig. 5. Microevolution led to decreased sensitivity of the Evo strain to different cell wall perturbing agents. (**A**) Resistance of analyzed strains against different cell wall stressors. The *cph1Δ/efg1Δ* strain was sensitive to all stresses while the Evo strain regained WT resistance (representative pictures of three experiments are shown). (**B**) Flow cytometry analysis of mannan and β-glucan exposure on the surface of live cells. Differences in fluorescence intensity between *cph1Δ/efg1Δ* strain and Evo strain point to an altered cell wall composition. Mean fluorescence intensity+SD of n = 3 (*$p<0.05$). (**C**) Western blot analysis to identify phosphorylated Mkc1 and Cek1 in *C. albicans* strains grown under non-stress conditions (control) or conditions of cell wall stress (450 µg/ml congo red) for 4 hours. Cell wall stress triggered phosphorylation of Mkc1 and Cek1 in the Evo strain, but not in the *cph1Δ/efg1Δ* strain. Tubulin served as loading control.
doi:10.1371/journal.pgen.1004824.g005

analyzed the expression levels of TF genes in the *cph1Δ/efg1Δ* and Evo strains under filament-inducing conditions in more depth (S2 DS7 Table). A significantly higher expression of 21 TF genes was shared by both strains, and only five TF genes were specifically upregulated in the *cph1Δ/efg1Δ* strain as compared to the levels in the Evo strain. Interestingly, 17 TF genes had significantly higher expression specifically in the Evo strain and not in *cph1Δ/efg1Δ*, including three genes known to be important hyphal morphogenesis regulators: *UME6* (in agreement with previous qRT-PCR results), *RIM101* and *HAC1*. Eight of the higher expressed TF genes in the Evo strain have unknown biological functions.

In the *cph1Δ/efg1Δ* strain, but not in the Evo strain, *CPH2*, *TEC1* and *ACE2*, which encode TFs involved in hyphal growth, were significantly downregulated under filament-inducing conditions. Finally, a significantly lower expression was observed for *NRG1* in the Evo strain, which codes for a repressor of hyphal development [42]. Hence, we scanned for Nrg1 binding sites $(A/C)(A/C/G)C_3T$ [43] in the putative promoter regions of genes specifically upregulated twofold in the Evo strain and detected the sequence motifs in 70% of these promoter regions (S2 DS8 Table). With this, the Nrg1 binding motif is statistically overrepresented in promoters of upregulated genes ($p<0.01$) when compared to promoters of all other genes. The downregulation of *NRG1* in the Evo strain may therefore facilitate expression of filament-associated genes and hence filament formation.

Further analyses indicated a significant upregulation of genes encoding for secreted aspartyl proteases (*SAP5*, *SAP6*, *SAP10*). In addition, significant differences in expression of genes associated with cell wall biosynthesis (*CHK1*, *KRE6*, *GLC3*, *MP65*, *ALG11* and *MNT2*), alkalinisation (*ACH1*) as well as of genes involved in glucose and galactose interconversion and uptake (*GAL10*, *GAL1*, *HGT2*, *HGT4*, *HGT12* and *GSY1*) were observed.

In summary, our transcriptional analysis indicated that serial passage through macrophages led to substantial alterations of the global transcriptional profile. The programs and pattern we found differed clearly from the *cph1Δ/efg1Δ* mutant, and resembled more the well-known programs of the wild type strain. This is concomitant with and likely correlated with the regained ability of the Evo strain to induce filaments and to induce damage to host cells *in vitro* and *in vivo*.

Comparative whole genome re-sequencing identifies mutations potentially linked to Cph1/Efg1-independent filamentation

We went on to determine the genetic basis for the observed phenotypical differences. No obvious large-scale structural variations were detectable between the karyotypes of wild type, the *cph1Δ/efg1Δ* and Evo strains using pulsed field gel electrophoresis (PFGE; S4A Figure). To detect possible loss of heterozygosity (LOH) events [44], we analyzed four SNP-restriction fragment length polymorphism (RFLP) markers per chromosome [45]. No differences were detected between double mutant and Evo strain (S3 DS1 Table). Taken together, these data show that no gross chromosomal rearrangements have occurred in the Evo strain.

We re-sequenced the genomes of the Evo and the *cph1Δ/efg1Δ* strains to identify single nucleotide polymorphisms (SNP) that may have arisen during the microevolution experiment. Sequencing depth for *cph1Δ/efg1Δ* and Evo were 99× and 108× in average, respectively, with 98.8% of the *C. albicans* SC5314 reference genome covered in both cases. Comparison of both sequences

revealed a chromosome 7 trisomy in the *cph1Δ/efg1Δ* strain, an aneuploidy that appears to have been lost during the evolution experiment (S4B Figure). This is also reflected by a 1.5× higher mean transcription level of genes on chromosome 7 in the *cph1Δ/efg1Δ* strain (S4C Figure). In addition, an amplification of *URA3* on chromosome 3 was observed. *URA3* was originally used as a marker to delete *CPH1* and *EFG1* in the *cph1Δ/efg1Δ* strain, and is now present in three copies in this mutant. The Evo strain contained 7–8 copies (S4B Figure). A qPCR analysis on isolated gDNA supported these findings (S4D Figure). PFGE and subsequent hybridization with a *URA3* specific probe further revealed that all copies were located on the same chromosome (S4E Figure). To exclude any possible contribution of multiple *URA3* gene copies to the filamentous phenotype, the Evo strain was cured from *URA3* with 5-fluoroorotic acid treatment [46]. This Evo Ura$^-$ strain was still able to filament, showing that *URA3* copy number is not responsible for the filamentous phenotype (S4F Figure). Additionally, after re-introduction of a single *URA3* using the standard CIp10 plasmid at the *RPS10* locus [47], these strains exhibited the same adhesion, invasion and macrophage damage properties as their multi-*URA3* counterparts (S4G Figure). This indicates that the excessive *URA3* copies do not have an influence on classical virulence properties of *C. albicans*.

We observed a high number of SNPs in the *cph1Δ/efg1Δ* strain: altogether, 70,197 heterozygous and 3,156 homozygous SNPs were identified in *cph1Δ/efg1Δ* relative to the *C. albicans* SC5314 consensus reference genome (Assembly 21, [48]). Similarly, 72,315 heterozygous and 3,294 homozygous SNPs were identified in the Evo strain. These figures are consistent with those achieved when reads obtained by sequencing the genome of *C. albicans* SC5314 are aligned on the reference genome and reflect the high level of heterozygosity in *C. albicans* as well as putative sequencing errors and ambiguous positions in the reference genome (homozygous SNPs). After combining these sets and filtering, only 329 putative SNPs were found to distinguish the *cph1Δ/efg1Δ* and Evo strains. Notably, polymorphisms at 209 of these positions are observed in the genomes of 19 clinical isolates, distributed over several *C. albicans* phylogenetic groups (CdE, unpublished data). This suggests that they were most likely not responsible for the restoration of filamentation. Of the 120 remaining positions, 83 were in non-coding regions, 22 resulted in synonymous changes and 15 resulted in non-synonymous changes (S3 DS2-4 Table). Finally, the RNA-Seq dataset was used as an additional source to detect SNPs specifically in expressed genes (see Materials and Methods & S3 DS6&7 Table,): A total of 65 putative transcribed SNPs, both heterozygous and homozygous, were found in the Evo strain, of which 21 were located in non-coding regions. Inside ORFs, 26 caused a synonymous and 13 a non-synonymous nucleotide exchange. Of all 39 SNPs detected in coding regions, 24 were located in genes of the *ALS* gene family (*ALS2* and *ALS4*), although these are likely false positives, as genes of the *ALS* family possess a very high sequence similarity and tandem repeat regions complicating read-mapping and SNP resolution [49]. Comparison of SNPs detected by RNA-Seq and Whole-Genome Sequencing revealed three SNPs shared by both detection methods. One SNP was located in a non-coding region between two uncharacterized genes (orf19.351 and orf19.352), while the other two were located inside ORFs. A SNP in *ATP18* (orf19.2066.1) resulted in a synonymous amino acid exchange, while the second SNP in *SSN3* (orf19.794) resulted in a heterozygous, non-synonymous Arg/Arg to Arg/Gln amino acid change.

A Mutation in *SSN3* is essential for the filamentous phenotype in the Evo strain

As the SNP at nucleotide position 1,055 in the *SSN3* ORF (Fig. 6A) was detected in both analyses, we focused our investigation on this specific mutation. Ssn3 has been well characterized in *Saccharomyces cerevisiae* as an RNA polymerase II holoenzyme-associated cyclin-dependent kinase of the Mediator complex contributing to transcriptional control [50]. It was shown that Ssn3 promotes the degradation of the transcription factor Ste12 by phosphorylation and thereby regulates *S. cerevisiae* filamentous growth [51]. As depicted in Fig. 6B, the heterozygous $Arg^{352}Gln$ mutation of Ssn3 in the Evo strain is located within the activation segment of the protein kinase catalytic domain. An amino acid sequence comparison of *C. albicans* Ssn3 to sequences from *S. cerevisiae*, *Cryptococcus neoformans*, *Mus musculus* and *Homo sapiens* demonstrated this arginine residue to be conserved from fungi to mammals. The activation segment comprises several conserved structural features: the magnesium binding loop, the activation loop and the P+1 loop, in which the mutation occurred. While the activation loop is the site of regulatory phosphorylation in many kinases, the P+1 loop forms a pocket that recognizes the substrate protein [52].

To ascertain the impact of the SNP on filamentation induction, we selectively deleted either the mutated or the wild type *SSN3* allele in the Evo strain, using the dominant selection marker *SAT1* [53]. Sanger sequencing confirmed the exclusive presence of either one allele in the genome (Fig. 7A). Strikingly, when incubated in DMEM with 10% serum at 37°C and 5% CO_2 only the strain with the mutated allele still present (Evo *ssn3Δ/SSN3$_m$*) was able to induce and maintain filamentation. The mutant containing only the wild type allele (Evo *SSN3/ssn3$_m$Δ*) remained in the elongated yeast form, and thus presented the typical ancestral (*cph1Δ/efg1Δ*) phenotype (Fig. 7A). In addition, only the Evo *ssn3Δ/SSN3$_m$* strain could escape from macrophages by forming filaments like the wild type (Fig. 7A). The damage capacity correlated with this ability to produce filaments: While Evo and Evo *ssn3Δ/SSN3$_m$* strains showed the same levels of phagocyte lysis, the Evo *SSN3/ssn3$_m$Δ* strain caused significantly less damage during co-incubation with macrophages. In fact, damage was indistinguishable from the original *cph1Δ/efg1Δ* strain (Fig. 7B). In contrast, the deletion of the mutated allele had no influence on the hyphal development defect on solid medium (S5A Figure) and sensitivity to cell wall disturbing agents (S5B Figure).

To further ascertain that the *SSN3* mutation alone is sufficient to allow filamentation in a *cph1Δ/efg1Δ* background, we created an independent *cph1Δ/efg1Δ* double mutant using the dominant selection marker *SAT1* (see Protocol S1). Importantly, this *cph1Δ/efg1Δ$_{SAT1}$* strain contained neither the *URA3* amplification nor the trisomy of chromosome 7 or other genetic alterations of the original *cph1Δ/efg1Δ* strain. In all our filamentation assays, this mutant behaved identical to the original *cph1Δ/efg1Δ* strain by not forming any hyphae (Fig. 7A) and hence not escaping from or damaging macrophages (Figs. 7A & 7B). To isolate the effect of the mutated *SSN3*, we followed several strategies with this new mutant: *SSN3* overexpression strains were created of both the wild type and mutated (*SSN3$_m$*) allele under the control of the strong *ADH1* promoter (see S1 Protocol). Strikingly, only the mutated allele allowed hyphae formation under inducing conditions in the *cph1Δ/efg1Δ$_{SAT1}$* strain (Fig. 7A, lower left corner), even in the presence of the two native *SSN3* alleles. Similarly, macrophage lysis was increased in the *SSN3$_m$* overexpressing strain, but not under *SSN3* overexpression (Fig. 7B, right panel). Finally, we integrated the mutated *SSN3* (together with a *SAT1* cassette) into the *SSN3* locus of *cph1Δ/efg1Δ$_{SAT1}$*, replacing one *SSN3* allele

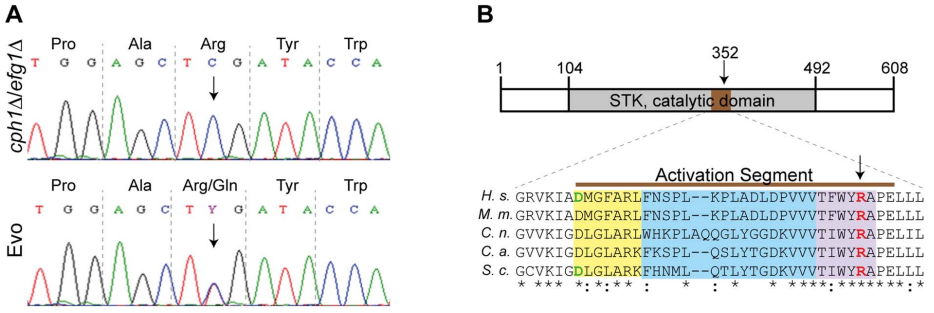

Fig. 6. Single nucleotide polymorphism in *SSN3* of the Evo strain and location of the mutated amino acid. (**A**) Partial *SSN3* sequence for *cph1Δ/efg1Δ* and Evo strains flanking SNP 1055 (marked with an arrow). Notice the heterozygosity in the Evo strain. (**B**) Schematic view of the catalytic domain of Ssn3 (STK = serine/threonine kinase) with the position of the activation segment highlighted in brown and the amino acid exchange indicated by an arrow (top). Sequence alignment of the Ssn3 activation segment in different species (*H. s. Homo sapiens* [NP_001251.1], *M. m. Mus musculus* [NP_705827.2], *C. n. Cryptococcus neoformans* [XP_568416.1], *C. a. C. albicans* [XP_720918.1] and *S. c. Saccharomyces cerevisiae* [NP_015283.1]). The arrow indicates the amino acid exchange in the Evo strain. The Mg-binding loop is highlighted in yellow, the activation loop in blue and the P+1 loop in purple. Amino acids that are known to abrogate kinase activity when mutated are colored in green [51,98]. Asterisks underneath the alignment indicate positions with conserved amino acids and colons indicate highly similar residues (bottom). The mutated arginine (red) is part of the highly conserved P+1 substrate recognition loop.
doi:10.1371/journal.pgen.1004824.g006

and essentially reproducing the heterozygous situation of the Evo strain. Again, this strain behaved virtually identical to the Evo strain, both in forming hyphae (Fig. 7A) and in damaging macrophages (Fig. 7B, right panel).

In summary, these data show that a non-synonymous mutation in *SSN3* that arose during our microevolution experiment is alone sufficient for regaining the ability to filament even in the absence of Efg1 and Cph1.

Discussion

Previous experimental studies on the acquisition of antifungal drug resistance and on stress-induced chromosome rearrangements have elegantly demonstrated the adaptive potential of *C. albicans* [44,54]. Here, we demonstrate – to our knowledge for the first time – that a complex trait such as the hyphal formation program of *C. albicans* can be subject to microevolution in the laboratory.

The yeast-to-hyphae transition is of crucial importance for full *C. albicans* pathogenicity, which is reflected by its complex regulation [14]. Multiple overlapping as well as separate signaling pathways are activated by various environmental signals to regulate hyphae formation. Wild type hyphae are an important contributor to the fungus' ability to escape from engulfing macrophages. In contrast, the *cph1Δ/efg1Δ* mutant strain cannot escape by filament formation, yet is able to replicate inside macrophages and to block phagosome maturation. Therefore, we expected that the mutant strain would survive in the phagosome, albeit with reduced fitness compared to the wild type. We monitored the phenotypic changes of the *cph1Δ/efg1Δ* strain co-passaged with macrophages for 42 passages. On a comparatively short evolutionary timescale our experiment resulted in a strain which not only regained the ability to filament, but also re-acquired other important characteristics, like a more wild type-like cell wall structure and increased virulence. We were able to show that a minimal sequence alteration accounts for the striking phenotypic reversal to wild type-like filamentation: a single missense mutation in *SSN3*. *SSN3* encodes a fungal protein kinase, which phosphorylates various regulators in *S. cerevisiae*. Our data shows that it can become important for bypassing the requirements of Efg1 and Cph1 for filamentation in *C. albicans*.

The in-depth characterization of the evolved strain revealed that the hyphal morphogenesis program can be induced by certain, but not all conditions which induce filamentation in the wild type strain. The fact that the Evo strain filaments in liquid, but not on solid media indicates an involvement of cAMP signaling and hence argues for a bypass of Efg1 functions rather than Cph1 [55–57]. This was further supported by three additional findings. First, the yeast-to-filament switch occurred in response to either serum, GlcNAc or CO_2, stimuli all known to trigger the activation of PKA signaling [23,58–60]. Second, filamentation was entirely blocked by the addition of the quorum-sensing molecule farnesol which represses both cAMP-PKA and MAPK signaling pathways [61,62]. The full restoration of filamentation when cAMP was added supports the involvement of the cAMP-PKA pathway. Third, the repressor of hyphae formation, Nrg1 is normally downregulated by the cAMP-PKA pathway, except in the presence of farnesol [63]. Transcriptome analysis showed *NRG1* expression to be downregulated in the Evo strain, but not in the *cph1Δ/efg1Δ* mutant. As 70% of the upregulated genes in the Evo strain contain an Nrg1 binding site, these data emphasize the likely importance of Nrg1 levels on filamentation of the Evo strain.

Given that the *cph1Δ/efg1Δ* mutant is strongly reduced in virulence [22], the almost wild type-level virulence in the Evo strain in our murine model was striking. Examination of kidney sections revealed filament formation of the Evo strain *in vivo*. Compared to wild type filaments, these were shorter and resulted in less pronounced tissue invasion, which is likely associated with the lower overall virulence compared to the wild type.

Three factors are likely to have contributed to the increased virulence of the Evo strain in the absence of Efg1 and Cph1: First, its ability to escape from macrophages like the wild type; second, its adhesion to host cells which was significantly higher than the *cph1Δ/efg1Δ* strain; and third, the ability to form filaments upon

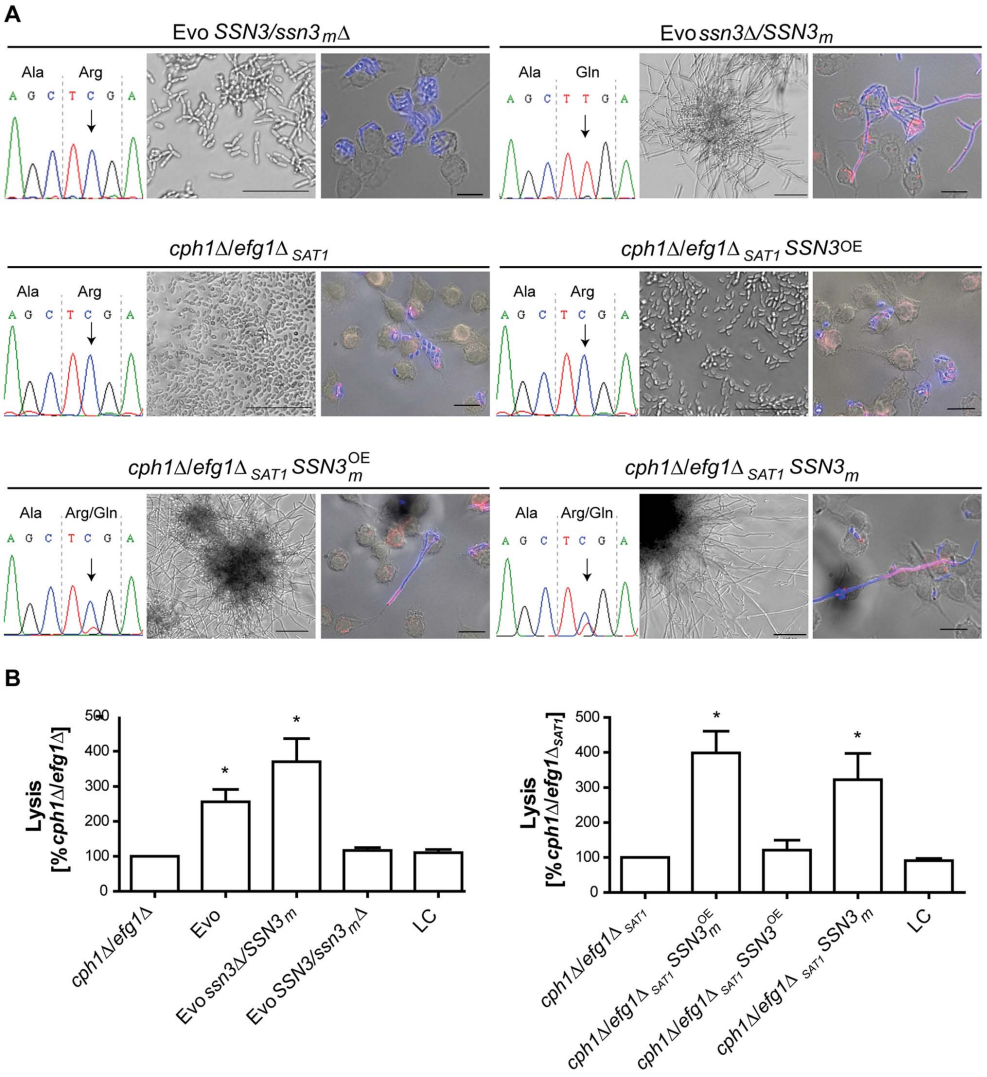

Fig. 7. A single nucleotide polymorphism in *SSN3* is essential for filamentation. (**A**) Distinct impact on morphology by: deleting either the mutated (*SSN3*/*ssn3*$_m$Δ) or the wild type *SSN3* allele (*ssn3*Δ/*SSN3*$_m$) in the Evo strain, by overexpressing either the wild type *SSN3* allele (*cph1*Δ/*efg1*Δ$_{SAT1}$*SSN3*^{OE}) or the mutated *SSN3* allele (*cph1*Δ/*efg1*Δ$_{SAT1}$*SSN3*$_m^{OE}$) or by expressing the mutated *SSN3* allele from its native locus (*cph1*Δ/*efg1*Δ$_{SAT1}$*SSN3*$_m$) in a newly generated *cph1*Δ/*efg1*Δ$_{SAT1}$ strain. The partial *SSN3* sequences demonstrate the homozygosity or heterozygosity of the *SSN3* allele (left). Filamentous growth is visible in the Evo *ssn3*Δ/*SSN3*$_m$, *cph1*Δ/*efg1*Δ$_{SAT1}$*SSN3*$_m^{OE}$ and *cph1*Δ/*efg1*Δ$_{SAT1}$*SSN3*$_m$ strains after growth for 18 h at 37°C and 5% CO_2 in DMEM+10% FBS and during co-incubation with macrophages, but not with the Evo *SSN3*/*ssn3*$_m$Δ strain, *cph1*Δ/*efg1*Δ$_{SAT1}$ and *cph1*Δ/*efg1*Δ$_{SAT1}$*SSN3*^{OE} strains (scale bars: 18 h, 50 μm and MΦ, 20 μm; representative pictures are shown) (right). (**B**) Cell damage of macrophages caused by the different strains, as determined by lactate dehydrogenase (LDH) assay after 32 h of co-incubation. Robust host cell damage depends on the presence of the mutated allele *SSN3m*. Mean and SD of n = 4 (*p<0.05; compared to *cph1*Δ/*efg1*Δ and *cph1*Δ/*efg1*Δ$_{SAT1}$ respectively; LC = low control, medium only).
doi:10.1371/journal.pgen.1004824.g007

contact with epithelial cells, which is a prerequisite for both active penetration into and induced endocytosis by host cells [64]. Wächtler *et al*. [17] showed that filamentation alone is insufficient to cause damage of host cells. We therefore compared the damage capacities of the *cph1*Δ/*efg1*Δ and the Evo strains. The Evo strain exhibited a significantly increased potential to damage both macrophages and epithelial cells compared to the double mutant. The adaptation to macrophages was accompanied by differences in additional traits, such as resistance to cell wall stresses. In the *cph1*Δ/*efg1*Δ strain, the higher sensitivity to cell wall disturbing agents, as well as the modified exposure of cell wall components, likely reflect an altered cell wall organization which was restored in the Evo strain. This is supported by findings from a recent study by Zavrel *et al*. [32] which showed that deletion of *EFG1* alone affects cell wall architecture. In our strains, these modifications of the cell wall seemed to be mediated by the kinases Mkc1 and Cek1. Previous analyses carried out in *cek1*Δ and *mkc1*Δ mutants already indicated their direct relationship to cell wall composition and integrity [35,65,66].

By analyzing the differences in gene expression acquired during co-culture passaging with macrophages, we found that all genes belonging to the core filamentation network [38] were upregulated in the Evo strain. This suggests that during filamentation the Evo strain transcriptionally utilizes the complete filamentation program. The transcription factors Tec1, Brg1, Ume6, Rim101, Hac1 and Efh1, which are known to be involved in regulation of filamentation [24,67–71], were also upregulated in the Evo strain. Together with Nrg1, they likely orchestrate filament formation in the Evo strain. For *UME6*, it has been shown that its transcription is repressed by Nrg1-Tup1 and that ectopic Ume6 expression in *cph1*Δ/*efg1*Δ can rescue the filamentation defect under certain conditions [69].

For the maintenance of hyphal extension, both *UME6* and *EED1* are central [27,31] and both showed an increased expression in the evolved strain. Thus, the mechanisms of hyphal extension seems similar between Evo and wild type cells [27]. Hence, the transcriptional conditions for initiation and maintenance of filamentation, which comprise the release of repression and the upregulation of positive regulators of filamentation, are met in the Evo strain. Furthermore, a considerable number of transcripts specifically up- and downregulated in the Evo strain are both *Candida*-specific and uncharacterized. It is feasible, therefore, that these uncharacterized transcripts assumed a novel role specifically during filament formation in the Evo strain. This is especially true as the morphological switch is one of the best-investigated characteristics in *C. albicans*, and genes involved in this process are generally well studied. However, differential regulation of genes not clearly linked to the yeast-to-hyphal switch, including these genes, but also *WOR1* and *NAT4* (both involved in the white-opaque switching) and *SST2* (involved in the mating response pathway), could have been caused by the mutated Ssn3 kinase (see below). Finally, it also should be noted that, even in the absence of filamentation, *cph1*Δ/*efg1*Δ was able to upregulate certain genes described as hyphae-associated under the condition tested here (incubation in DMEM+10% FBS at 37°C and 5% CO_2 on a plastic surface). This is in disagreement with previous data showing that *EFG1* is required for expression of several hyphae-associated genes [23,72]. It is possible, however, that alternative pathway(s), such as the Rim101 pH response pathway, are involved, as the cells were simultaneously exposed to diverse stimuli for filamentation. However, these genes still showed an increased induction in the Evo strain compared to the *cph1*Δ/*efg1*Δ strain, which argues for a further adaptation-induced, filament-associated change in regulation.

It has been demonstrated that acquired drug resistance in *C. albicans* is often accompanied by aneuploidy and/or isochromosome formation [54,73] and that several stress conditions can enhance the rates of LOH events likely by mitotic recombination [44]. However, we did not detect any LOH events between the *cph1*Δ/*efg1*Δ and the Evo strain. The chromosome 7 trisomy was present initially in the *cph1*Δ/*efg1*Δ strain [74] and the Evo strain restored disomy by loss of one copy. The remaining gross genetic difference, an *URA3* amplification in the Evo strain can be explained by an insufficient Ura3 expression from the *EFG1* locus. An amplification of the gene may have increased fungal fitness during our experiment by ensuring more transcripts and hence more efficient growth. Prior studies suggested that ectopic expression of *URA3* influences the phenotypes of a diverse range of mutants [75,76], and duplication of a *hisG-URA3-hisG* cassette resulted in restored filamentation of an *hwp1*Δ mutant [77]. We were able to exclude these Ura3 effects as causes for the Evo strain filamentation, as the acquired filamentation phenotype was maintained after removal of the multiple *URA3* copies. Furthermore, after re-introduction of a single copy of *URA3*, no differences in virulence traits, like adhesion or invasion, were detectable as compared to the multi-copy strains. Overall, these data and the fact that the observed filamentation and other phenotypes persisted even after repassaging in rich (YPD) medium, argued for small-scale genomic alterations, rather than epigenetic changes, acquired by *cph1*Δ/*efg1*Δ cells adapting to macrophages.

Comparative genome sequencing (by WGS and RNA-Seq) of *cph1*Δ/*efg1*Δ and Evo strains allowed us to pinpoint the microevolutionary changes in the Evo strain at the single nucleotide level. By combining the different approaches, we detected an expressed SNP in *SSN3*, which resulted in an Arg-to-Gln change at a highly conserved position within the presumable protein kinase domain. This SNP and thus gain of heterozygosity was found to be central for the yeast-to-filament transition of the Evo strain. Deletion of the mutated *SSN3* allele prevented the morphological switch in the Evo strain during growth under filament-inducing conditions and interaction with several types of host cells. Other phenotypes specific to the Evo strain were not affected by the deletion of the mutated *SSN3* allele, suggesting that they evolved independently from filamentation.

Importantly, introduction of a single mutated allele into an independent *efg1*Δ/*cph1*Δ strain fully copied the filamentation and host cell damage phenotype of the Evo strain. This strain contained neither the multiple *URA3* copies nor the trisomy of chromosome 7 or any other possible genetic alterations of the original *efg1*Δ/*cph1*Δ strain. Hence, the *SSN3* mutation alone bypassed the lack of the central transcription factors Cph1 and Efg1 and restored the ability to cause host cell damage *in vitro*, and likely to induce higher virulence *in vivo*.

Ssn3 itself (also referred as Srb10 or Cdk8) is part of the CDK (cyclin-dependent kinase) module (SRB10/11) of the Mediator complex, which is a regulator of RNA-polymerase II (RNAP II) activity [78,79]. This CDK module phosphorylates the largest subunit of RNAP II, and Ssn3 additionally has roles in both transcriptional activation and repression in response to physiological signals, coordinating gene expression. By regulating the stability of the two important regulators, Ste12 (ortholog of Cph1) and Phd1, Ssn3 in *S. cerevisiae* is involved in the differentiation of yeasts into pseudohyphae under nutrient-limiting conditions [51,80]. Interestingly, a kinase-deficient Asp290Ala Ssn3 only weakly phosphorylates Ste12 *in vitro*, and the lack of phosphorylation increases its stability [51]. Moreover, the catalytic activity of Ssn3 contributes to the repression of a subset of Tup1-regulated

genes [81–83] in *S. cerevisiae*. Tup1 is recruited to promoters by Nrg1 [84], a factor which was downregulated in the Evo strain.

Although the precise signaling pathway(s) controlling Ssn3 remain to be determined, Chang *et al.* [85] showed that the activity of Srb9, another subunit of the CDK kinase module, is regulated by the PKA signaling pathway in *S. cerevisiae*. Based on our data it is tempting to speculate that the activation of the cAMP-PKA pathway results in activation of Ssn3 kinase activity, and the observed filament-specific transcriptional changes may thus depend on either a reduced or absent substrate recognition or on impaired substrate phosphorylation activity due to the $Arg^{352}Gln$ substitution. It is supposable that a loss of the substrate-specific kinase activity increases the stability of positive regulator(s) of filamentation by reducing their phosphorylation. Alternatively (or in addition), the impaired kinase activity could lead to a derepression of genes associated with positive regulation of filamentous growth. Importantly, in this model the kinase-deficient Ssn3 remains part of the Mediator complex, and could fulfill any additional function it may have (e.g. in the structure or recruitment of additional proteins). In both models, a decreased kinase activity would reduce inhibitory effects on filamentation, and hence would increase the sensitivity of the filamentation network to external stimuli. This would likely allow to bypass the need for additional Efg1 signaling. Additional genes outside of the immediate filamentation network may also be affected, as this model implies a pleiotropic effect of the Ssn3 mutation, with several transcription factors as possible clients.

Thus, it seems that not the disrupted cAMP-PKA signaling pathway itself evolved in our microevolution experiment, but instead a regulatory hub for filamentation which the pathway probably targets in addition to Efg1. In this hub, even single or few mutations seem to be able to lead to striking phenotypic alterations, as many filament-associated genes are directly or indirectly targeted. Finally, it is interesting to speculate why only one *SSN3* allele was mutated, and we did not observe any LOH event to homozygosity at this locus. It seems possible that one mutated allele alone was sufficient to promote filamentation in macrophages, while the other wild type allele, still capable of full phosphorylation activity, was still required for additional functions of Ssn3. This is somewhat supported by the observation that overexpression of the mutated *SSN3* allele in a background with the native *SSN3* alleles still in place was sufficient to allow hyphae formation. In our model, the mutated Ssn3 competes with the wild type Ssn3, and overexpression allows the mutated protein to gain entry into a sufficient number of Mediator complexes.

In conclusion, using the nonfilamentous mutant *cph1*Δ/*efg1*Δ, we have shown that *C. albicans* can rescue one of its key virulence traits, the yeast-to-hyphal switch, with a single nucleotide change when put under adequate selection pressure. A mutation in the transcriptional regulator Ssn3 adaptively rewired the transcription network to enable filamentation in response to external cues while bypassing the need for Efg1 and Cph1. This shows an unexpected robustness of the whole filamentation system even to severe disruptions, and a high degree of adaptability. The selection scenario we used - co-incubation with macrophages - clearly reflects a condition *C. albicans* encounters in the host and thus might be an evolutionary pressure that can shape the infection biology of this fungus. In fact, this hypothesis is supported by another evolution experiment, which analyzed the adaptation of *C. glabrata* to macrophages. There, the selection pressure resulted in the appearance of a strain with pseudohyphae-like structures and increased virulence again by a single nucleotide mutation [86]. This demonstrates that during interaction with the host or host cells, significant changes in morphology and virulence are possible on a very short evolutionary time-scale.

Materials and Methods

Ethic statement

All animal experiments were in compliance with the German animal protection law and were approved by the responsible Federal State authority (Thüringer Landesamt für Lebensmittelsicherheit und Verbraucherschutz) and ethics committee (beratende Kommission nach § 15 Abs. 1 Tierschutzgesetz; permit no. 03-007/07).

Body surface temperature and body weight were recorded daily and animals were monitored twice a day for disease progression. Mice showing severe signs of illness (isolation from the group, apathy, hypothermia and drastic weight loss) were humanely sacrificed by ketamine/xylazine overdose and exsanguination.

Strains and growth conditions

Candida albicans strains and mutants used in this study are listed in S1 Table. Strains were grown in YPD medium (1% peptone, 1% yeast extract, 2% glucose and optionally 2% agar) or SD medium (2% dextrose, 0.17% yeast nitrogen base, 0.5% ammonium sulfate and optionally 2% agar) at 30°C. Uridine (50 μg/ml) or nourseothricin (NAT; 100 μg/ml) were added as required. If not stated otherwise, stationary phase cells were used in the experiments. Mutants were constructed as described in Protocol S1.

Cell lines

The murine peritoneal macrophage-like cell line J774A.1 (DSMZ) and the human buccal carcinoma epithelial cell line TR-146 (Cancer Research Technology) were grown in Dulbecco's Modified Eagle's Medium (DMEM, PAA) supplemented with 10% FBS (PAA) and routinely cultured until passage 20. Both cell lines were maintained at 37°C under 5% CO_2. J774A.1 cells were removed from tissue-culture flasks by gentle scraping, while TR-146 cells were enzymatically harvested by Accutase (PAA) treatment.

Evolution experiment

About 8×10^6 J774A.1 macrophages were seeded into a 75 cm^2 cell culture flask with DMEM supplemented with 10% FBS and 1% Penicillin/Streptomycin (PAA). For the evolution experiment, macrophages were initially infected with 4×10^6 cells of the *cph1*Δ/*efg1*Δ strain. After that, 4×10^6 re-isolated *C. albicans* cells were transferred to a fresh macrophage culture. After 24 h of co-incubation, infected macrophages were washed (3× with PBS) and lysed with 2 ml lysis buffer (50 mM Tris, 5 mM EDTA, 150 mM NaCl and 0.5% Nonidet P40 [Sigma-Aldrich]). The lysate was transferred to a 2 ml reaction tube and fungal cells were collected by centrifugation. The *C. albicans* cells were washed two times with DMEM and counted before infection of fresh macrophages.

To verify the absence of *EFG1* and *CPH1*, Southern blot analysis was performed for the Evo strain as described previously [22]. Briefly, genomic DNA (gDNA) was digested with *Ava*II or *Kpn*I to verify *EFG1* or *CPH1* deletion, respectively. DIG-labeled probes were generated (Roche) using genomic DNA from the strain SC5314 and primers EFG-A/EFG-B and P33/CPH-B (S1 Table).

Phenotypic characterization

A detailed description of the phenotypic analyses can be found in Protocol S1.

Staining procedures and detection of β-1,3-glucans and mannans

Fungal cells were grown in DMEM+10% FBS on glass coverslips in a 24 well microtiter plate for filipin (Sigma), calcofluor white (CFW) and Als3 immunostaining. Flow cytometry was used to quantify mannan and β-1,3-glucan exposure on the surface of stationary *C. albicans* cells after staining with concanavalin A and anti- β-1,3-glucan. Piercing and invasion rate were determined by differential staining. All staining procedures are described in Protocol S1. Epifluorescence (Leica DM5500B, Leica DFC360 FX) was used to detect CFW and filipin (DAPI filter), Alexa Fluor 488 (FITC filter) and Alexa Fluor 647 (Cy5 filter). Micrographs were taken with a Leica Digital Camera DFC360 FX or a Zeiss AxiCam ICc3.

Replication and piercing assay

Two times 10^5 J774A.1 macrophages were seeded onto glass cover slips placed in 24 well microtiter plates and allowed to adhere overnight. Non-adherent macrophages were removed by washing with PBS. To monitor intracellular replication, *C. albicans* cells were labeled with 100 µg/ml fluorescein isothiocyanate (FITC, Sigma-Aldrich) in carbonate buffer (0.1 M Na_2CO_3, 0.15 M NaCl, pH 9.0) for 30 min at 37°C and washed 3× with PBS. To quantify piercing rates, cells were washed without prior staining. Two times 10^5 fungal cells were added to macrophages in DMEM+10% FBS. The plates were incubated for indicated timepoints (see figure legends). Cells were then washed once with PBS and fixed with 4% paraformaldehyde. Intracellular replication was detected by fluorescence microscopy after mounting the samples in ProLong Gold Antifade Reagent with DAPI (Invitrogen). After co-incubation, piercing of macrophages by filaments was quantified by differential staining. The assays were performed in biological triplicates.

Adherence and invasion assay

Two times 10^5 TR-146 epithelial cells were seeded onto glass cover slips placed in 24 well microtiter plates and cultured for 2–3 days to 95%-100% confluency. Adherence and invasion assays were performed as previously described [17]. Briefly, to determine the adhesion rate, TR-146 monolayers were infected with 1×10^6 *C. albicans* cells. After one hour of co-incubation, non-adherent yeast cells were removed by rinsing 3× with PBS. Cells were fixed with 4% paraformaldehyde, permeabilized with 0.5% Triton X-100 and adherent *C. albicans* cells were stained with CFW for fluorescence microscopy. Invasion rates were determined by infecting TR-146 monolayers with 1×10^5 *C. albicans* cells. After incubation, cells were fixed and differentially stained for fluorescence microscopy. Both assays were repeated at least three times.

Quantification of damage to host cells

Five times 10^4 host cells (J774A.1 or TR-146) were seeded in 96 well microtiter plates. J774A.1 macrophages were cultured for 1 day before use, while TR-146 epithelial cells were cultured for 2 days to 95%–100% confluency. Damage of macrophages and epithelial cells was determined by measuring the release of lactate dehydrogenase (LDH) with the Cytotoxicity Detection Kit (Roche Applied Science) following 32 h of co-incubation with 5×10^4 *C. albicans* cells according to the manufacturer's protocol. The experiments were performed as previously described [17] and repeated at least three times.

Murine infection model

For survival studies the intravenous challenge model for disseminated *C. albicans* infection was used. Six to eight weeks old female BALB/c mice (18–20 g) purchased from Charles River were used for the experiments. Mice were challenged intravenously with 5×10^5 *C. albicans* cells in 200 µl PBS via the lateral tail vein. All mice surviving to day 20 were humanely sacrificed. For histology, kidneys were collected and fixed with buffered formalin and paraffin-embedded sections were stained with Periodic acid-Schiff (PAS) according to standard protocols.

Western blot analysis

To detect phosphorylated Mkc1 and Cek1 as well as α-tubulin, cells of an overnight culture were adjusted to an OD_{600} of 0.5 in SD medium (control) or SD medium supplemented with 450 µg/ml congo red, and incubated for 4 hours at 30°C. Cell disruption, protein extraction and western blot analysis using anti-phospho-p44/42 MAP kinase antibody (Cell Signalling Technology) and rat anti-α-tubulin antibody (AbD Serotec), respectively, were performed as previously described [87].

RNA sample preparation and isolation

C. albicans cells from an overnight culture were diluted to $OD_{600} = 0.2$ in YPD medium and grown to log-phase for 4 h at 30°C. Cells were collected by centrifugation and a zero time point sample was frozen in liquid nitrogen until RNA extraction. In addition, 1×10^7 cells were incubated one hour under filament-inducing conditions (DMEM+10% FBS at 37°C and 5% CO_2 in a 75 cm^2 cell culture flask). For farnesol experiments, 10 µM farnesol was added to the medium just prior to the experiment. After incubation, medium and non-adherent cells were removed and 5 ml ice-cold PBS was added. The cells were collected by scraping and then centrifuged for 5 min at 6,000 g at 4°C. Cell pellets were snap frozen in liquid nitrogen. Total RNA was isolated using the Ribopure-Yeast Kit (Ambion) and treated with Turbo DNase (Ambion). RNA quality was determined in a Bioanalyzer with an RNA 6000 Nano LabChip Kit (Agilent Technologies) according to the manufacturer's protocol. RNA concentration was determined with a Nanodrop ND1000 (Peqlab).

Copy number determination and quantitative gene expression analysis

Copy number and expression levels of selected genes were analyzed with a my-Budget 5× EvaGreen QPCR Mix II (Bio&Sell) in a C1000TM Thermal Cycler (BioRad) using gene-specific primers (S1 Table). For expression analysis, 600 ng of total RNA was reversely transcribed with the SuperScript III First-Strand Synthesis Kit (Invitrogen) according to the manufacturer's instructions. *URA3* gene copy number was determined from 100 ng of gDNA with primers URA3-fw and URA3-re (S1 Table). PCR conditions were as followed: 95°C for 15 min, 40 cycles of each 95°C for 15 s, 60°C for 40 s and 72°C for 15 s. A melting profile was generated to confirm PCR product specificity. Relative gene expression levels were determined by the $2^{\Delta\Delta Ct}$ method [88] with *ACT1*, *EFB1* and *PMA1* as internal controls. *URA3* copy number was calculated with *ACT1* internal control and gDNA from SC5314 (containing two copies of *URA3*) as reference. Three independent experiments were performed.

Pulsed-field gel electrophoresis (PFGE) and SNP-RFLP analysis

PFGE and SNP-RFLP are described in Protocol S1.

RNA sequencing, transcriptional profiling and SNP discovery from RNA-Seq data

In order to use only high quality reads, trimming was performed using Btrim (window size = 15, average quality score = 20) [89].

For differential gene expression analysis high quality trimmed reads were mapped against the sequence Assembly 21 of strain SC5314 [48] using the spliced read mapper TopHat 2.0.6 [90] with the "known transcripts" (-G option) and uniquely mapped reads were counted using HTSeq [86]. Raw counts for each gene were loaded into R and differentially expressed genes were identified using the packages edgeR and DESeq [91,92] and filtered by adjusted p-values (<0.01) and RPKM value (≥ 1). Data were deposited at the Gene Expression Omnibus (GSE56174) and can be found in S2 Table. Nrg1 binding sites (A/C)(A/C/G)C_3T in putative promoter regions (usually −1000 bp/+50 bp) of all *C. albicans* genes were determined by SiTaR [93] allowing no mismatch. Fisher's exact test was used to determine if the Nrg1 motif-containing promoters were overrepresented in genes specifically upregulated twofold in the Evo strain, as compared to all remaining genes. For SNP calling quality trimmed reads from all samples of each strain were merged and the protocol of GATK [94] with slight changes was followed (i.e. reads were mapped using BWA algorithm [95], duplicates were removed and realignment around indels and base recalibration was performed). Next, we used bam-readcount (www.github.com/genome/bam-readcount), which determines the nucleotide distribution at each single base. Heterozygous SNPs were defined as positions where 25% or more of the reads showed an alternative nucleotide. Homozygous SNPs were defined as positions where more than 90% of the reads differed from the reference. Minimum nucleotide sequence depth was 20. Clustal Omega [96] was used for multiple sequence alignments.

Whole genome sequencing and sequence analysis

Genomic DNA isolated from the *cph1*Δ/*efg1*Δ and Evo strains were processed to prepare libraries for Illumina sequencing, and the TruSeq DNA Sample Prep kit (Illumina) was used according to the manufacturer's recommendations. DNAs were randomly fragmented by sonication to an average fragment length of 500 bp and Illumina adapters were blunt-end ligated to the fragments. The final libraries were amplified by PCR followed by sequencing on an Illumina Genome Analyzer platform (Illumina GAII). 60 nt single-end reads were aligned to the *C. albicans* strain SC5314 reference genome [48] downloaded on 02/24/2012 using shore 5.0 [97]. Sequencing depth scores were computed for each 1 kb region across the genomes and for ORFs using sequencing depth data for each nucleotide located within the 1 kb region or the ORF. Sequencing depth scores were normalized based on the overall sequencing depth obtained for each genome. Single nucleotide polymorphisms were identified using shore 5.0 [97] at positions covered at least 30 times with a minimum quality of 25. Homozygous SNPs were defined as positions where 90% of the reads meeting these criteria differed from the reference genome. Heterozygous SNPs were defined as positions where 20% or more of the reads showed one allele and 80% or less of the reads showed a second allele.

Statistical analysis

Data were visualized and statistically analyzed using GraphPad Prism version 5.00 (GraphPad Software, USA). Statistical analyses were performed by 1-way ANOVA (mannan and β-1,3-glucan exposure) or 2-way ANOVA (piercing, adhesion, invasion, damage and gene expression) followed by a Bonferroni correction. Differences in survival of mice were evaluated by Log-rank (Mantel-Cox) test.

Supporting Information

S1 Figure Screening of *C. albicans* deletion mutants for defects in hyphal formation during interaction with macrophages and verification of the *cph1*Δ/*efg1*Δ genotype in the Evo strain. (**A**) Morphologies of different *C. albicans* mutants and the corresponding wild type strains upon phagocytosis by macrophages (J774A.1). Note that only the double mutant strain cannot escape from macrophages. Figures show overlay of DIC and fluorescent images. *C. albicans* appears blue (CFW) and extracellular section of hypha red (ConA). Arrows highlight piercing of macrophage membrane by *C. albicans* cells (scale bar: 10 μm). (**B**) The WT (CAI4+CIp10) forms hyphae after phagocytosis (white arrows), while the *cph1*Δ/*efg1*Δ strain replicates intracellularly in J774A.1 cells (black arrows). Yeast cells were stained with FITC prior to infection. Six hours after infection samples were fixed and stained with DAPI for analysis by fluorescence microscopy. FITC is not transferred to hypha or new daughter cells during cell division (scale bar: 10 μm, representative picture). (**C**) Southern blots of wild type (WT), *cph1*Δ/*efg1*Δ and Evo strains confirm the deletion of *EFG1* (left) and *CPH1* (right) in the Evo strain. Genomic DNAs were digested with either *Ava*II or *Kpn*I, and DNA molecular weight marker III, DIG labeled (Roche) was used as size standard.
(TIF)

S2 Figure Morphological index and morphology of the Evo strain under different conditions. (**A**) Overnight cultures of *cph1*Δ/*efg1*Δ and Evo strains were diluted into DMEM+10% FBS and incubated for the indicated time points at 37°C and 5% CO_2 on cover slips. After fixation the percentage of yeast, pseudohyphal and hyphal cells was quantified using the morphological index (MI) [26]. Mean+SD of at least 100 cells in two experiments. (**B**) Morphology of *cph1*Δ/*efg1*Δ and Evo strains in phase contrast under different filament-inducing conditions. The Evo strain formed filaments in response to stimuli other than used during the evolution experiment. (**C**) Colony morphology of strains under embedded conditions. The Evo strain exceeds the hyperfilamentation phenotype of the *cph1*Δ/*efg1*Δ strain. (**D**) Colony morphology of analyzed strains grown on solid YPD agar supplemented with 10% FBS for 6 days, on solid Spider and Lee's medium for 9 days, and on solid YNB agar supplemented with 2% glucose and 10 mM urea for 11 days at 37°C (scale bar: 1 mm; representative pictures are shown). Only WT forms filaments under these conditions.
(TIF)

S3 Figure Results of the RNA-Seq analysis. (**A**) RNA-Seq results are in good agreement with qRT-PCR analyses of four selected genes. Expression was normalized against three housekeeping genes (*ACT1*, *EFB1* and *PMA1*). Fold change gene expression of filament-inducing condition versus yeast promoting condition (YPD, 30°C) is shown (mean+SD). (**B**) Venn diagrams of differentially expressed genes of *cph1*Δ/*efg1*Δ and Evo strains during growth in DMEM+10% FBS at 37°C and 5% CO_2 on a plastic surface compared to yeast promoting condition (YPD, 30°C). (**C**) Expression heat map of the eight genes of the core filamentation response in the *cph1*Δ/*efg1*Δ and Evo strains during growth in DMEM+10% FBS at 37°C and 5% CO_2 with white boxes indicating no differential expression (NDE).
(TIF)

S4 Figure Analysis of the genome architecture of the Evo strain. (**A**) Whole genome profiles of wild type (WT), *cph1*Δ/*efg1*Δ and Evo chromosomes separated by PFGE and stained with ethidium bromide are identical among all strains analyzed. (**B**) Whole-genome sequencing of *cph1*Δ/*efg1*Δ and Evo strains revealed a chromosome 7 (Chr 7) trisomy in the *cph1*Δ/*efg1*Δ strain (top) and an amplification of *URA3* on chromosome 3 (Chr 3) in the Evo strain (bottom). Normalized $\log_2$ read depth per 1 kb region along the chromosomes are shown. (**C**) Relative expression from

the RNA-Seq experiment between *cph1*Δ/*efg1*Δ and Evo. The chromosome 7 trisomy of *cph1*Δ/*efg1*Δ is reflected by a mean 1.5× higher expression of genes specifically on this chromosome. Other chromosomes showed no change in mean expression (shown here for chromosome 6 as an example) (**D**) Verification of *URA3* amplification by qPCR analysis using genomic DNA isolated from the *cph1*Δ/*efg1*Δ and Evo strains. Copy number of *URA3* was normalized to *ACT1*. Mean+SD. (**E**) The *URA3* gene was successfully deleted from the Evo Ura$^-$ strain. Whole chromosomes of the Evo strain and 5-fluoroorotic acid (FOA) treated Evo cells (Evo Ura$^-$) were separated by PFGE, stained with ethidium bromide (left) and subjected to Southern hybridization with a *URA3* probe to verify successful deletion of all *URA3* copies in the genome (right). (**F**) Phenotype of Evo and Evo Ura$^-$ strains after growth for 18 h at 37°C and 5% CO_2 in DMEM+10% FBS (representative pictures). Removal of all *URA3* copies has no influence on the regained filamentation properties of the Evo strain. (**G**) Virulence traits of the *efg1*Δ/*cph1*Δ (ΔΔ) and Evo strains cured of *URA3* (Ura-) and after subsequent re-introduction of a single *URA3* copy using the CIp10 plasmid for genomic integration. No difference can be detected between the multi-copy and single-copy strains in adhesion or invasion of epithelial cells, or damage to macrophages. Absence of *URA3* reduces macrophage damage, probably due to the low uridine concentration inside the phagosome.
(TIF)

S5 Figure Characterization of the Evo *SSN3*/*ssn3*$_m$Δ and Evo *ssn3*Δ/*SSN3*$_m$ strains. (**A**) Colony morphology of the Evo *ssn3*Δ/*SSN3*$_m$ strain grown on solid Lee's medium, Spider medium and YPD medium supplemented with 10% FBS for 6 days at 37°C (scale bar: 1 mm; representative pictures are shown). Observed morphologies resemble those of the Evo strain. (**B**) Stress resistance of *cph1*Δ/*efg1*Δ, Evo, Evo *SSN3*/*ssn3*$_m$Δ and Evo *ssn3*Δ/*SSN3*$_m$ strains against different cell wall perturbing agents. The presence of either *SSN3* allele has no influence on the increased stress resistance of the Evo strain. Experiments yielded similar results for at least three replicates.
(TIF)

S1 Table Strains and primers used in the study.
(DOC)

S2 Table Excel data sheets contain data for differentially expressed genes in the *cph1*Δ/*efg1*Δ and the Evo strains during incubation under filament-inducing condition (DMEM+10% FBS at 37°C and 5% CO_2 on a plastic surface) identified by RNA-Seq analysis (log_2 RPKM data and raw read counts for all annotated transcripts and novel transcripts identified by Bruno *et al.* [37]).
(XLS)

S3 Table SNP-RFLP analysis of 32 SNP-RFLP markers (two per chromosome arm) to detect large-scale changes in the genomes of the *cph1*Δ/*efg1*Δ and Evo strains and predicted SNPs discovered by whole-genome sequencing and RNA-Seq of *cph1*Δ/*efg1*Δ and Evo strains.
(XLS)

S1 Protocol Additional methods used in this manuscript and for generation of the supplementary figures.
(DOC)

Acknowledgments

We thank G. Fink for providing the original *efg1*Δ/*cph1*Δ mutant. We thank U. Stöckel and B. Weber for their assistance with the mouse infection experiment, B. Hebecker and J. Voigt for their help with flow cytometry, and N. Jablonowski and D. Schulz for their invaluable assistance with the experiments. We also thank M. von der Heide for help with PFGE and S. G. Filler for kindly providing the Als3 antibody. We are grateful to L. Kasper, B. Hebecker, K. Seider, M. Polke and D. Wilson for their suggestions and helpful discussions.

Author Contributions

Conceived and designed the experiments: AW RM IDJ KK OK AF Cd SB BH. Performed the experiments: AW RM MS IDJ SJ AF Cd. Analyzed the data: AW JL RM FH IDJ SJ TW AF Cd SB. Contributed reagents/materials/analysis tools: IDJ AF RG OK Cd SB BH. Wrote the paper: AW RM SJ AF Cd SB RG BH.

References

1. Pfaller MA, Diekema DJ (2007) Epidemiology of invasive candidiasis: a persistent public health problem. Clin Microbiol Rev 20: 133–163.
2. Kumamoto CA, Vinces MD (2005) Contributions of hyphae and hypha-co-regulated genes to *Candida albicans* virulence. Cell Microbiol 7: 1546–1554.
3. Ene IV, Brunke S, Brown AJ, Hube B (2014) Metabolism in Fungal Pathogenesis. Cold Spring Harb Perspect Med.
4. Kvaal CA, Srikantha T, Soll DR (1997) Misexpression of the white-phase-specific gene *WH11* in the opaque phase of *Candida albicans* affects switching and virulence. Infect Immun 65: 4468–4475.
5. Mayer FL, Wilson D, Hube B (2013) *Candida albicans* pathogenicity mechanisms. Virulence 4: 119–128.
6. van de Veerdonk FL, Plantinga TS, Hoischen A, Smeekens SP, Joosten LA, et al. (2011) *STAT1* mutations in autosomal dominant chronic mucocutaneous candidiasis. N Engl J Med 365: 54–61.
7. van Enckevort FH, Netea MG, Hermus AR, Sweep CG, Meis JF, et al. (1999) Increased susceptibility to systemic candidiasis in interleukin-6 deficient mice. Med Mycol 37: 419–426.
8. White TC, Pfaller MA, Rinaldi MG, Smith J, Redding SW (1997) Stable azole drug resistance associated with a substrain of *Candida albicans* from an HIV-infected patient. Oral Dis 3 Suppl 1: S102–109.
9. Rustchenko-Bulgac EP, Sherman F, Hicks JB (1990) Chromosomal rearrangements associated with morphological mutants provide a means for genetic variation of *Candida albicans*. J Bacteriol 172: 1276–1283.
10. Shin JH, Park MR, Song JW, Shin DH, Jung SI, et al. (2004) Microevolution of *Candida albicans* strains during catheter-related candidemia. J Clin Microbiol 42: 4025–4031.
11. Tsong AE, Tuch BB, Li H, Johnson AD (2006) Evolution of alternative transcriptional circuits with identical logic. Nature 443: 415–420.
12. Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD (2008) The evolution of combinatorial gene regulation in fungi. PLoS Biol 6: e38.
13. Forche A, Magee PT, Selmecki A, Berman J, May G (2009) Evolution in *Candida albicans* populations during a single passage through a mouse host. Genetics 182: 799–811.
14. Sudbery PE (2011) Growth of *Candida albicans* hyphae. Nat Rev Microbiol 9: 737–748.
15. Jacobsen ID, Wilson D, Wächtler B, Brunke S, Naglik JR, et al. (2012) *Candida albicans* dimorphism as a therapeutic target. Expert Rev Anti Infect Ther 10: 85–93.
16. Filler SG, Sheppard DC (2006) Fungal invasion of normally non-phagocytic host cells. PLoS Pathog 2: e129.
17. Wächtler B, Wilson D, Haedicke K, Dalle F, Hube B (2011) From attachment to damage: defined genes of *Candida albicans* mediate adhesion, invasion and damage during interaction with oral epithelial cells. PLoS One 6: e17046.
18. Wellington M, Koselny K, Sutterwala FS, Krysan DJ (2014) *Candida albicans* triggers NLRP3-mediated pyroptosis in macrophages. Eukaryot Cell 13: 329–340.
19. Uwamahoro N, Verma-Gaur J, Shen HH, Qu Y, Lewis R, et al. (2014) The pathogen *Candida albicans* hijacks pyroptosis for escape from macrophages. MBio 5: e00003–00014.
20. Lorenz MC, Bender JA, Fink GR (2004) Transcriptional response of *Candida albicans* upon internalization by macrophages. Eukaryot Cell 3: 1076–1087.
21. Gow NA, van de Veerdonk FL, Brown AJ, Netea MG (2011) *Candida albicans* morphogenesis and host defence: discriminating invasion from colonization. Nat Rev Microbiol 10: 112–122.
22. Lo HJ, Köhler JR, DiDomenico B, Loebenberg D, Cacciapuoti A, et al. (1997) Nonfilamentous *C. albicans* mutants are avirulent. Cell 90: 939–949.
23. Stoldt VR, Sonneborn A, Leuker CE, Ernst JF (1997) Efg1p, an essential regulator of morphogenesis of the human pathogen *Candida albicans*, is a member of a conserved class of bHLH proteins regulating morphogenetic processes in fungi. Embo J 16: 1982–1991.

24. Doedt T, Krishnamurthy S, Bockmühl DP, Tebarth B, Stempel C, et al. (2004) APSES proteins regulate morphogenesis and metabolism in *Candida albicans*. Mol Biol Cell 15: 3167–3180.
25. Martin SW, Konopka JB (2004) Lipid raft polarization contributes to hyphal growth in *Candida albicans*. Eukaryot Cell 3: 675–684.
26. Merson-Davies LA, Odds FC (1989) A morphology index for characterization of cell shape in *Candida albicans*. J Gen Microbiol 135: 3143–3152.
27. Martin R, Moran GP, Jacobsen ID, Heyken A, Domey J, et al. (2011) The *Candida albicans*-specific gene *EED1* encodes a key regulator of hyphal extension. PLoS One 6: e18394.
28. Zhao X, Oh SH, Cheng G, Green CB, Nuessen JA, et al. (2004) *ALS3* and *ALS8* represent a single locus that encodes a *Candida albicans* adhesin; functional comparisons between Als3p and Als1p. Microbiology 150: 2415–2428.
29. Nobile CJ, Fox EP, Nett JE, Sorrells TR, Mitrovich QM, et al. (2012) A recently evolved transcriptional network controls biofilm development in *Candida albicans*. Cell 148: 126–138.
30. Brown DH, Jr., Giusani AD, Chen X, Kumamoto CA (1999) Filamentous growth of *Candida albicans* in response to physical environmental cues and its regulation by the unique *CZF1* gene. Mol Microbiol 34: 651–662.
31. Banerjee M, Thompson DS, Lazzell A, Carlisle PL, Pierce C, et al. (2008) *UME6*, a novel filament-specific regulator of *Candida albicans* hyphal extension and virulence. Mol Biol Cell 19: 1354–1365.
32. Zavrel M, Majer O, Kuchler K, Rupp S (2012) Transcription factor Efg1 shows a haploinsufficiency phenotype in modulating the cell wall architecture and immunogenicity of *Candida albicans*. Eukaryot Cell 11: 129–140.
33. Gow NA, Hube B (2012) Importance of the *Candida albicans* cell wall during commensalism and infection. Curr Opin Microbiol 15: 406–412.
34. Eisman B, Alonso-Monge R, Roman E, Arana D, Nombela C, et al. (2006) The Cek1 and Hog1 mitogen-activated protein kinases play complementary roles in cell wall biogenesis and chlamydospore formation in the fungal pathogen *Candida albicans*. Eukaryot Cell 5: 347–358.
35. Navarro-Garcia F, Alonso-Monge R, Rico H, Pla J, Sentandreu R, et al. (1998) A role for the MAP kinase gene *MKC1* in cell wall construction and morphological transitions in *Candida albicans*. Microbiology 144 (Pt 2): 411–424.
36. Roman E, Nombela C, Pla J (2005) The Sho1 adaptor protein links oxidative stress to morphogenesis and cell wall biosynthesis in the fungal pathogen *Candida albicans*. Mol Cell Biol 25: 10611–10627.
37. Bruno VM, Wang Z, Marjani SL, Euskirchen GM, Martin J, et al. (2010) Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. Genome Res 20: 1451–1458.
38. Martin R, Albrecht-Eckardt D, Brunke S, Hube B, Hünniger K, et al. (2013) A core filamentation response network in *Candida albicans* is restricted to eight genes. PLoS One 8: e58613.
39. Pierce JV, Dignard D, Whiteway M, Kumamoto CA (2013) Normal adaptation of *Candida albicans* to the murine gastrointestinal tract requires Efg1p-dependent regulation of metabolic and host defense genes. Eukaryot Cell 12: 37–49.
40. Samaranayake YH, Cheung BP, Yau JY, Yeung SK, Samaranayake LP (2013) Human serum promotes *Candida albicans* biofilm growth and virulence gene expression on silicone biomaterial. PLoS One 8: e62902.
41. Hope H, Schmauch C, Arkowitz RA, Bassilana M (2010) The *Candida albicans* ELMO homologue functions together with Rac1 and Dck1, upstream of the MAP Kinase Cek1, in invasive filamentous growth. Mol Microbiol 76: 1572–1590.
42. Braun BR, Kadosh D, Johnson AD (2001) *NRG1*, a repressor of filamentous growth in *C.albicans*, is down-regulated during filament induction. Embo J 20: 4753–4761.
43. Murad AM, Leng P, Straffon M, Wishart J, Macaskill S, et al. (2001) *NRG1* represses yeast-hypha morphogenesis and hypha-specific gene expression in *Candida albicans*. Embo J 20: 4742–4752.
44. Forche A, Abbey D, Pisithkul T, Weinzierl MA, Ringstrom T, et al. (2011) Stress alters rates and types of loss of heterozygosity in *Candida albicans*. MBio 2.
45. Forche A, Steinbach M, Berman J (2009) Efficient and rapid identification of *Candida albicans* allelic status using SNP-RFLP. FEMS Yeast Res 9: 1061–1069.
46. Fonzi WA, Irwin MY (1993) Isogenic strain construction and gene mapping in *Candida albicans*. Genetics 134: 717–728.
47. Murad AM, Lee PR, Broadbent ID, Barelle CJ, Brown AJ (2000) CIp10, an efficient and convenient integrating vector for *Candida albicans*. Yeast 16: 325–327.
48. Inglis DO, Arnaud MB, Binkley J, Shah P, Skrzypek MS, et al. (2012) The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. Nucleic Acids Res 40: D667–674.
49. Hoyer LL (2001) The *ALS* gene family of *Candida albicans*. Trends Microbiol 9: 176–180.
50. Kuchin S, Yeghiayan P, Carlson M (1995) Cyclin-dependent protein kinase and cyclin homologs *SSN3* and *SSN8* contribute to transcriptional control in yeast. Proc Natl Acad Sci U S A 92: 4006–4010.
51. Nelson C, Goto S, Lund K, Hung W, Sadowski I (2003) Srb10/Cdk8 regulates yeast filamentous growth by phosphorylating the transcription factor Ste12. Nature 421: 187–190.
52. Nolen B, Taylor S, Ghosh G (2004) Regulation of protein kinases; controlling activity through activation segment conformation. Mol Cell 15: 661–675.
53. Reuss O, Vik A, Kolter R, Morschhauser J (2004) The *SAT1* flipper, an optimized tool for gene disruption in *Candida albicans*. Gene 341: 119–127.
54. Selmecki AM, Dulmage K, Cowen LE, Anderson JB, Berman J (2009) Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. PLoS Genet 5: e1000705.
55. Liu H, Köhler J, Fink GR (1994) Suppression of hyphal formation in *Candida albicans* by mutation of a *STE12* homolog. Science 266: 1723–1726.
56. Bockmühl DP, Ernst JF (2001) A potential phosphorylation site for an A-type kinase in the Efg1 regulator protein contributes to hyphal morphogenesis of *Candida albicans*. Genetics 157: 1523–1530.
57. Leberer E, Harcus D, Broadbent ID, Clark KL, Dignard D, et al. (1996) Signal transduction through homologs of the Ste20p and Ste7p protein kinases can trigger hyphal formation in the pathogenic fungus *Candida albicans*. Proc Natl Acad Sci U S A 93: 13217–13222.
58. Jung WH, Stateva LI (2003) The cAMP phosphodiesterase encoded by *CaPDE2* is required for hyphal development in *Candida albicans*. Microbiology 149: 2961–2976.
59. Castilla R, Passeron S, Cantore ML (1998) N-acetyl-D-glucosamine induces germination in *Candida albicans* through a mechanism sensitive to inhibitors of cAMP-dependent protein kinase. Cell Signal 10: 713–719.
60. Klengel T, Liang WJ, Chaloupka J, Ruoff C, Schröppel K, et al. (2005) Fungal adenylyl cyclase integrates CO_2 sensing with cAMP signaling and virulence. Curr Biol 15: 2021–2026.
61. Roman E, Alonso-Monge R, Gong Q, Li D, Calderone R, et al. (2009) The Cek1 MAPK is a short-lived protein regulated by quorum sensing in the fungal pathogen *Candida albicans*. FEMS Yeast Res 9: 942–955.
62. Davis-Hanna A, Piispanen AE, Stateva LI, Hogan DA (2008) Farnesol and dodecanol effects on the *Candida albicans* Ras1-cAMP signalling pathway and the regulation of morphogenesis. Mol Microbiol 67: 47–62.
63. Lu Y, Su C, Wang A, Liu H (2011) Hyphal development in *Candida albicans* requires two temporally linked changes in promoter chromatin for initiation and maintenance. PLoS Biol 9: e1001105.
64. Zakikhany K, Naglik JR, Schmidt-Westhausen A, Holland G, Schaller M, et al. (2007) *In vivo* transcript profiling of *Candida albicans* identifies a gene essential for interepithelial dissemination. Cell Microbiol 9: 2938–2954.
65. Navarro-Garcia F, Eisman B, Fiuza SM, Nombela C, Pla J (2005) The MAP kinase Mkc1p is activated under different stress conditions in *Candida albicans*. Microbiology 151: 2737–2749.
66. Galán-Diez M, Arana DM, Serrano-Gomez D, Kremer L, Casasnovas JM, et al. (2010) *Candida albicans* beta-glucan exposure is controlled by the fungal *CEK1*-mediated mitogen-activated protein kinase pathway that modulates immune responses triggered through dectin-1. Infect Immun 78: 1426–1436.
67. Schweizer A, Rupp S, Taylor BN, Rollinghoff M, Schröppel K (2000) The TEA/ATTS transcription factor CaTec1p regulates hyphal development and virulence in *Candida albicans*. Mol Microbiol 38: 435–445.
68. Cleary IA, Lazzell AL, Monteagudo C, Thomas DP, Saville SP (2012) *BRG1* and *NRG1* form a novel feedback circuit regulating *Candida albicans* hypha formation and virulence. Mol Microbiol 85: 557–573.
69. Zeidler U, Lettner T, Lassnig C, Müller M, Lajko R, et al. (2009) *UME6* is a crucial downstream target of other transcriptional regulators of true hyphal development in *Candida albicans*. FEMS Yeast Res 9: 126–142.
70. Davis D, Wilson RB, Mitchell AP (2000) *RIM101*-dependent and-independent pathways govern pH responses in *Candida albicans*. Mol Cell Biol 20: 971–978.
71. Wimalasena TT, Enjalbert B, Guillemette T, Plumridge A, Budge S, et al. (2008) Impact of the unfolded protein response upon genome-wide expression patterns, and the role of Hac1 in the polarized growth, of *Candida albicans*. Fungal Genet Biol 45: 1235–1247.
72. Sharkey LL, McNemar MD, Saporito-Irwin SM, Sypherd PS, Fonzi WA (1999) *HWP1* functions in the morphological development of *Candida albicans* downstream of *EFG1*, *TUP1*, and *RBF1*. J Bacteriol 181: 5273–5279.
73. Selmecki A, Gerami-Nejad M, Paulson C, Forche A, Berman J (2008) An isochromosome confers drug resistance *in vivo* by amplification of two genes, *ERG11* and *TAC1*. Mol Microbiol 68: 624–641.
74. Arbour M, Epp E, Hogues H, Sellam A, Lacroix C, et al. (2009) Widespread occurrence of chromosomal aneuploidy following the routine production of *Candida albicans* mutants. FEMS Yeast Res 9: 1070–1077.
75. Cheng S, Nguyen MH, Zhang Z, Jia H, Handfield M, et al. (2003) Evaluation of the roles of four *Candida albicans* genes in virulence by using gene disruption strains that express *URA3* from the native locus. Infect Immun 71: 6101–6103.
76. Sundstrom P, Cutler JE, Staab JF (2002) Reevaluation of the role of *HWP1* in systemic candidiasis by use of *Candida albicans* strains with selectable marker *URA3* targeted to the *ENO1* locus. Infect Immun 70: 3281–3283.
77. Sharkey LL, Liao WL, Ghosh AK, Fonzi WA (2005) Flanking direct repeats of hisG alter *URA3* marker expression at the *HWP1* locus of *Candida albicans*. Microbiology 151: 1061–1071.
78. Myers LC, Kornberg RD (2000) Mediator of transcriptional regulation. Annu Rev Biochem 69: 729–749.
79. Lewis BA, Reinberg D (2003) The mediator coactivator complex: functional and physical roles in transcriptional regulation. J Cell Sci 116: 3667–3675.
80. Raithatha S, Su TC, Lourenco P, Goto S, Sadowski I (2012) Cdk8 regulates stability of the transcription factor Phd1 to control pseudohyphal differentiation of *Saccharomyces cerevisiae*. Mol Cell Biol 32: 664–674.

81. Schüller J, Lehming N (2003) The cyclin in the RNA polymerase holoenzyme is a target for the transcriptional repressor Tup1p in *Saccharomyces cerevisiae*. J Mol Microbiol Biotechnol 5: 199–205.
82. Green SR, Johnson AD (2004) Promoter-dependent roles for the Srb10 cyclin-dependent kinase and the Hda1 deacetylase in Tup1-mediated repression in *Saccharomyces cerevisiae*. Mol Biol Cell 15: 4191–4202.
83. Kuchin S, Carlson M (1998) Functional relationships of Srb10-Srb11 kinase, carboxy-terminal domain kinase CTDK-I, and transcriptional corepressor Ssn6-Tup1. Mol Cell Biol 18: 1163–1171.
84. Park SH, Koh SS, Chun JH, Hwang HJ, Kang HS (1999) Nrg1 is a transcriptional repressor for glucose repression of *STA1* gene expression in *Saccharomyces cerevisiae*. Mol Cell Biol 19: 2044–2050.
85. Chang YW, Howard SC, Herman PK (2004) The Ras/PKA signaling pathway directly targets the Srb9 protein, a component of the general RNA polymerase II transcription apparatus. Mol Cell 15: 107–116.
86. Brunke S, Seider K, Fischer D, Jacobsen I, Kasper L, et al. (2014) Brunke S, Seider K, Fischer D, Jacobsen ID, Kasper L, et al. (2014) One Small Step for a Yeast - Microevolution within Macrophages Renders Candida glabrata Hypervirulent Due to a Single Point Mutation. PLoS Pathog 10(10): e1004478. doi:10.1371/journal.ppat.1004478.
87. Mayer FL, Wilson D, Jacobsen ID, Miramón P, Grosse K, et al. (2012) The novel *Candida albicans* transporter Dur31 Is a multi-stage pathogenicity factor. PLoS Pathog 8: e1002592.
88. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta CT}$ Method. Methods 25: 402–408.
89. Kong Y (2011) Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics 98: 152–153.
90. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111.
91. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140.
92. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.
93. Fazius E, Shelest V, Shelest E (2011) SiTaR: a novel tool for transcription factor binding site prediction. Bioinformatics 27: 2806–2811.
94. Van der Auwera GA, Mauricio OC, Hatl C, Poplin R, del Angel G, et al. (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 11.10.1–11.10.33.
95. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
96. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7: 539.
97. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, et al. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res 18: 2024–2033.
98. Akoulitchev S, Chuikov S, Reinberg D (2000) TFIIH is negatively regulated by cdk8-containing mediator complexes. Nature 407: 102–106.

2.4 Manuskript 3: »CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes«

Status veröffentlicht, Dezember 2015

Literaturangabe Wolf, Thomas ; Shelest, Vladimir ; Nath, Neetika ; Shelest, Ekaterina: CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. In: *Bioinformatics* (2015), Dezember, btv713. `http://dx.doi.org/10.1093/bioinformatics/btv713`. – DOI 10.1093/bioinformatics/btv713. – ISSN 1367–4803, 1460–2059

Übersicht Sekundärmetabolite sind von hohem pharmazeutischen und medizinischen Interesse. Gene des Sekundärmetabolismus sind häufig in Clustern organisiert, das heißt kolokalisiert und koreguliert. In dieser Arbeit stellen wir zwei neu entwickelte Computerprogramme vor: SMIPS (»Secondary Metabolites by InterProScan«) zur genomweiten Suche von Ankergenen des Sekundärmetabolismus und CASSIS (»Cluster Assignment by Islands of Sites«) zur Vorhersage von Sekundärmetabolit-Gen-Clustern. SMIPS basiert auf der funktionellen Annotation von Proteindomänen. CASSIS sucht in der Umgebung der von SMIPS gefundenen Ankergene nach Inseln, die mit cluster-spezifischen Sequenzmotiven angereichert sind. Beide Methoden wurden einer Kreuzvalidierung unterzogen und mit den Ergebnissen anderer bekannter Programme zur Cluster-Vorhersage verglichen.

Beiträge TW, VS und ES konzipierten und entwickelten die CASSIS-Methode. TW und ES konzipierten und entwickelten die SMIPS-Methode. TW sammelte die Trainingsdaten für SMIPS und CASSIS, implementierte beide Methoden und war an der Entwicklung der Internetseiten beteiligt. NN sammelte Daten für den Vergleich zwischen CASSIS/SMIPS und den anderen Programmen zur Cluster-Vorhersage, entwickelte die Internetseiten für CASSIS und SMIPS, und trug zur Verbesserung des Manuskriptes bei. TW und ES schrieben das Manuskript. TW war für das Layout und die Finalisierung des Manuskriptes zuständig.

Bioinformatics, 32(8), 2016, 1138–1143
doi: 10.1093/bioinformatics/btv713
Advance Access Publication Date: 9 December 2015
Original Paper

OXFORD

Sequence analysis

CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes

Thomas Wolf*, Vladimir Shelest, Neetika Nath and Ekaterina Shelest*

Research Group Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute (HKI), Jena 07745, Germany

*To whom correspondence should be addressed

Associate Editor: Alfonso Valencia

Received on 29 June 2015; revised on 3 November 2015; accepted on 28 November 2015

Abstract

Motivation: Secondary metabolites (SM) are structurally diverse natural products of high pharmaceutical importance. Genes involved in their biosynthesis are often organized in clusters, i.e., are co-localized and co-expressed. *In silico* cluster prediction in eukaryotic genomes remains problematic mainly due to the high variability of the clusters' content and lack of other distinguishing sequence features.

Results: We present Cluster Assignment by Islands of Sites (CASSIS), a method for SM cluster prediction in eukaryotic genomes, and Secondary Metabolites by InterProScan (SMIPS), a tool for genome-wide detection of SM key enzymes ('anchor' genes): polyketide synthases, non-ribosomal peptide synthetases and dimethylallyl tryptophan synthases. Unlike other tools based on protein similarity, CASSIS exploits the idea of co-regulation of the cluster genes, which assumes the existence of common regulatory patterns in the cluster promoters. The method searches for 'islands' of enriched cluster-specific motifs in the vicinity of anchor genes. It was validated in a series of cross-validation experiments and showed high sensitivity and specificity.

Availability and implementation: CASSIS and SMIPS are freely available at https://sbi.hki-jena.de/cassis.

Contact: thomas.wolf@leibniz-hki.de or ekaterina.shelest@leibniz-hki.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Secondary metabolites (SM), also often referred as natural products, are substances with outstanding diversity of biological activities, including pharmaceutically important ones, e.g. antibiotic, toxic, immunosuppressant. They are produced primarily by microorganisms (fungi, bacteria, algae). Genes responsible for SM biosynthesis and also for modifications, transport, regulation, etc., are often organized in clusters (Brakhage and Schroeckh, 2011). Here, we define clusters as sets of co-localized and co-regulated genes, the products of which are presumably functionally connected. In fungi, SM clusters typically have modest sizes (normally up to 20 genes), are characterized by tight co-localization of successive genes and are often regulated by a cluster-specific transcription factor (csTF), which can be a part of the respective cluster (Brakhage, 2013; Keller and Hohn, 1997). In many cases, also not csTF can regulate SM clusters (Hoffmeister and Keller, 2007). Recently, an example of cross-cluster regulation was described in fungi: activation of the csTF of a cluster led to upregulation of another cluster on a different chromosome; in addition to the own cluster (Bergmann *et al.*, 2010). In this example, elucidation of the cluster specific motif helped to understand the mode of regulation of the second cluster.

Two SM classes of particular importance are synthesized by multimodular megasynthases: polyketide synthases (PKS) and non-ribosomal peptide synthetases (NRPS) or PKS–NRPS hybrids. In eukaryotes, in particular in fungi, these enzymes are characterized by specific multidomain structure and large size, which makes them easy to detect in the genomes. The other cluster members, however, are more difficult to identify since the clusters' content varies greatly and there are no stable cluster markers (i.e. genes that would always accompany a megasynthase). This constitutes the first challenge for computational prediction of clusters.

The second challenge is the scarcity of the experimental data. The main body of experimental evidence for SMs and their biosynthetic pathways comes from bacteria and is not always applicable to eukaryotes. For instance, amino acid specificity of adenylation domains is quite well predictable from NRPS structure (Eppelmann *et al.*, 2002; Stachelhaus *et al.*, 1999) in bacteria but the same models do not work for fungi (Boettger *et al.*, 2012). Fungal data are rather scanty, in total <40 clusters are fully described so far (collected in this study, see Supplementary Table S1).

Most cluster prediction tools developed heretofore depend on domain homology, e.g. antiSMASH (Blin *et al.*, 2013), SMURF (Khaldi *et al.*, 2010), CLUSEAN (Weber *et al.*, 2009) or ClustScan (Starcevic *et al.*, 2008). These tools rely on collections of protein domains found in known clusters and predict new clusters by searching for these domains. This approach works well for similar clusters but has difficulties when encountering new cluster members (i.e. the proteins with new functions, with domains unknown to the system). Besides, it is known that not all successive genes in a cluster region belong to the cluster, e.g. at least four genes within the aflatoxin cluster are 'gap' genes that are not conserved and not assigned to the aflatoxin or sterigmatocystin biosynthesis (Amaike and Keller, 2011). Consideration of domains of the gap genes leads to erroneous predictions. All these problems together with the limited number of eukaryotic 'template' clusters make similarity-based methods error-prone and tending to overestimate the clusters' lengths, when applied to eukaryotes. Homology limitations might be bypassed by applying other approaches, such as window-averaged DNA curvature profiles (Do and Miyano, 2008) or methods relying on expression data, like microarrays, etc. (Andersen *et al.*, 2013). But these methods are limited in their applications. The former is restricted to LaeA-like regulated clusters, the latter require expression data, which can be problematic because most fungal clusters are silent under laboratory conditions (Brakhage and Schroeckh, 2011) and their induction is a challenging task.

Of all existing cluster predicting tools, antiSMASH is the most prominent, reliable and very much recommendable to use. Nonetheless, there is one type of useful information that is ignored by the similarity approach that is utilized by antiSMASH: the information about common TF binding sites that characterize the clusters. Since the cluster genes are co-regulated, their promoters should share the transcription factor binding sites (TFBS) for the common regulator. Taking into account this additional layer of information can improve the cluster prediction and supply with additional useful characteristics, such as the shared regulatory pattern and the nature of the regulating csTF.

Recently, we suggested an approach to detect eukaryotic gene clusters by estimating the density of binding motifs for csTF. The density must be higher within the clusters and lower, although not completely abolished, in other parts of the genome. The method, and the tool based on the method, is called Motif Density Method (MDM, Wolf *et al.*, 2013). MDM showed high specificity and sensitivity and was able to solve difficult problems like distinguishing closely located clusters (separated by just several genes), the task unsolvable for similarity-based tools (Wolf *et al.*, 2013). After having solved the main problem—the usage of promoter information for cluster prediction—we wanted to improve the method making the algorithm more transparent and the tool easier to handle.

Here, we present 'Cluster Assignment by Islands of Sites' (CASSIS), the further development and improvement of MDM. We made several changes, most importantly in the prediction algorithm, which are described in detail in the 'Methods' section. In short, instead of estimating the motifs' density in a sliding window, we applied a set of rules to identify the borders of the motif 'islands' around the anchor gene. The introduced changes improved the performance and made the algorithm simpler and more straightforward. The CASSIS method is implemented in a tool with the same name. An online version as well as downloads for Windows and Linux is available. Besides, we added a small tool called 'Secondary Metabolites by InterProScan' (SMIPS) for the fast and easy genome-wide detection of SM anchor genes, e.g. PKS, NRPS and dimethylallyl tryptophan synthases (DMATS). SMIPS results can be directly sent to the CASSIS tool or used separately to describe the SM biosynthetic potential of a species.

2 Methods

The SMIPS and CASSIS tools are two discrete software tools, with the option to run CASSIS on the output of SMIPS. In this section, we provide a step-by-step description of the entire work-flow.

2.1 Training data

A positive training set of 38 known (experimentally proven) SM gene clusters was used to estimate the parameters of the CASSIS search. This collection is restricted to clusters which have been verified experimentally by gene inactivation (disruption, deletion or knock-out), gene over-expression experiments, assigning gene functions to steps in the biosynthesis, or observable co-regulation of transcription. This set was manually collected based on literature and can be found in Supplementary Table S1.

For comparison with SMURF and antiSMASH, the training set for CASSIS included the 24 clusters that were published in 2010 or earlier (and hence could have been used for training of SMURF and antiSMASH, too). Whereas, the test set contained only the 12 clusters that were 'new' for all three compared tools, i.e. published in 2011 or later (see Supplementary Table S1).

The genome sequences, protein sequences and corresponding gene annotations were downloaded from the Broad Institute (http://www.broadinstitute.org) or Aspergillus Genome Database (Cerqueira *et al.*, 2014).

2.2 Evaluation

To assess the accuracy, precision, etc., of CASSIS and compare these characteristics with other tools, we ran cross-validation experiments. In each prediction, correctly identified cluster genes were considered as true positives (TP). The total number of TP was the sum of all genes of the considered clusters. The cluster genes not detected as TP by a predictive tool were counted as false negatives (FN), genes predicted outside the genuine clusters were false positives (FP). To obtain a feasible number of true negatives (TN), which are in general all non-cluster genes and hence make a huge number for a whole contig or a chromosome, we restricted the considered genomic region to ±30 genes around the anchor gene (because the largest so far known cluster—aflatoxin—contains about 30 genes). Note that this

restriction was used only for counting TN. For quantitative comparison of the tools, we calculated sensitivity (recall), specificity, precision, false discovery rate (FDR), accuracy and F_1-score according to standard definitions as derived from confusion matrix.

2.3 SMIPS tool

SMIPS is a small tool for the genome-wide prediction of PKS, NRPS and DMATS. SMIPS' input can be a protein FASTA file or an InterProScan output file. For the latter, SMIPS accepts the formats of InterProScan or the JGI tables (for details see https://sbi.hki-jena.de/smips/Help.php#Input).

The representative set of protein domain models for each enzyme type was collected by scanning available known fungal PKS, NRPS and DMATS. The collection was extended with known sets of SM domains from bacteria and plants (e.g. Blin *et al.*, 2013). The final list of all considered InterPro (IPR) accession numbers is shown in Supplementary Table S2.

SMIPS extracts all genes with at least one IPR number coinciding with the preselected SM domain list (Supplementary Table S2). The selected genes are evaluated for the occurrence of a set of domains typically sufficient for the full enzymatic activity ('minimal set' of domains characteristic for each SM type): KS, AT and ACP for PKS; A, C and T for NRPS; a single domain with prenyltransferase activity is sufficient for a DMATS (see Supplementary Table S2 for abbreviations). Incomplete NRPS and PKS forms (i.e. possessing more than one domain but not the minimal set) are reported as 'NRPS-like' or 'PKS-like'. Single KS, C and AT domains are reported separately (e.g. 'KS-only'). Finally, the domain arrangement of a gene is reported in simple text format (e.g. KS-AT-ACP). SMIPS output contains tables with all genes with at least one typical SM domain, and with all information for each putative SM gene: name, type, domain arrangement, etc.

SMIPS is very fast. On an Intel Core2Duo CPU, running at 3 GHz, it always takes less than a second to analyze the InterProScan files.

2.4 Choosing the promoter length

To estimate the optimal length of promoter sequences to be extracted, we performed an analysis of experimentally proven fungal TFBS from two databases (TRANSFAC, status of 2012, and FunTF, an in-house database for fungal TFBS). All TFBS were mapped on the respective genomic sequences and the distance to the corresponding transcription start site (TSS) was measured. The mapping results (Fig. 1) suggest that the overwhelming majority of genuine sites are located in the region −1000/+50 bp around the TSS. This range is therefore the recommendable length of promoter sequences, at least for the analysis of TFBS occurrences.

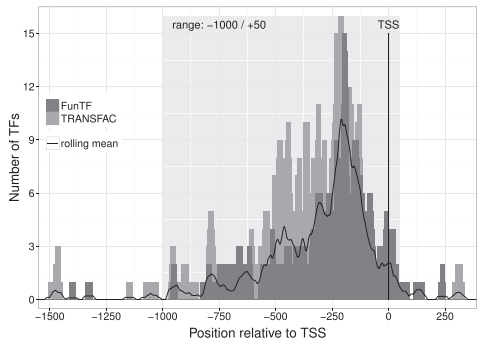

Fig. 1. Choosing the promoter range. The great majority of the genuine fungal TFBS from TRANSFAC and FunTF map to the region −1000/+50 bp

2.5 CASSIS tool for SM cluster predictions

CASSIS is the successor of MDM published in 2013 (Wolf *et al.*, 2013). It underwent several changes in the algorithm but the main idea remained the same: the sites for a TF regulating co-expressed cluster genes must be more dense (or form 'islands') within the cluster region.

CASSIS requires two input files: (i) genome sequence (contigs, chromosomes) in FASTA format; (ii) the corresponding annotation with start position, stop position and strand orientation of each gene. The user also needs a list of genes serving as 'anchors' for the future clusters. The latter can be SMIPS output or any other list of genes (e.g. manually selected). Principally, CASSIS is not restricted to only SM cluster predictions and will work for any anchor gene.

2.5.1 Promoter sequences

Before starting any prediction, CASSIS retrieves all promoter sequences genome-wide (based on the annotation file). The standard promoter range (−1000/+50 around TSS) applies if the intergenic region is >1 kb (or 2 kb for two non-overlapping promoters). If the promoter is bidirectional (overlapping) or the intergenic region is <1 kb, the whole intergenic region is retrieved. No promoter sequences are considered for genes overlapping by the 5′-ends.

2.5.2 Motif search

The tools MEME and FIMO (Bailey and Elkan, 1994; Grant *et al.* 2011; Bailey *et al.* 2009 (suite); http://www.meme-suite.org), required for the next two steps, are not incorporated into CASSIS and should be therefore pre-installed on the system.

The first three steps of the prediction (selection of the interim promoter sets, MEME and FIMO searches) are made as described in the initial MDM publication (Wolf *et al.*, 2013). In short, motifs (putative binding sites) are searched in interim sets of promoters around the anchor gene. Since the length of the cluster and the location of the anchor gene within the cluster are unknown, CASSIS automatically prepares several promoter sets around the anchor ranging from three to 15 promoters upstream and downstream the anchor gene, in total up to 250 different sets (Fig. 2). All sets are sent to MEME for prediction of over-represented motifs.

MEME is run for each set with the following search parameters: any number of repetitions (ANR); one motif to find; motif width 6–12 bp. To select the motifs for further analysis, CASSIS applies the following restrictions: (i) the motif must be found in the promoter of the anchor gene; (ii) the motif must be in more than one promoter; and (iii) the MEME *E*-value must not exceed a certain estimated cut-off (see Section 2.5.5). All MEME input and output files are preserved.

The motifs fulfilling the requirements are automatically sent to FIMO (Grant *et al.*, 2011), which predicts the motifs' occurrences in all promoters of the considered genome. Thus, the FIMO input is the FASTA file with genome-wide extracted promoter sequences. The search is restricted by a *p*-value cut-off (see Section 2.5.5). Based on the FIMO results, CASSIS counts the number of motifs per promoter. At this step, the motif can be rejected if: (i) it is not found in the promoter of the anchor gene (this can happen because of the FIMO cut-off); (ii) it is not found in any other but the anchor promoter; (iii) the motif is too frequent, i.e. is found in more than a

certain percentage of all promoter sequences (see Section 2.5.5 for parameter settings).

2.5.3 Transforming the genomic sequence into the sequence of promoters

On this step, the genomic sequence is seen as a sequence of promoters. This means that promoters are considered as units characterized by the number of occurrences of the considered motif (Fig. 2, Step D). The genomic sequence is in this way transformed into a string of numbers, each number representing the motifs' occurrences in a unit (promoter). For instance, if one motif was found in the first promoter, two motifs in the second, and 0 in the third and fourth promoters, the string will be 1–2–0–0. SM clusters should represent the regions with the highest density of the motifs (in other words, 'islands' of non-zero numbers in the number string).

2.5.4 Defining the cluster borders

The anchor gene's promoter is taken as seed for the cluster prediction. CASSIS scans the number string immediately upstream and downstream of the anchor promoter until it hits the first 'zero' value (promoter without binding site). If one or two zeroes are followed by a non-zero value, they are included in the cluster (gap rule '≤2 zero-promoters', see Section 2.5.5). If more than two zeroes are found in a row, the cluster is interrupted. The last non-gap promoter marks the border of the cluster prediction. This step is carried out for each motif (Section 2.5.2). If this leads to multiple different cluster borders, the most abundant one will be considered. The output of CASSIS contains the locus IDs of the first and last genes corresponding to the promoters flanking the predicted cluster, the motif and promoter information, and the length of each prediction.

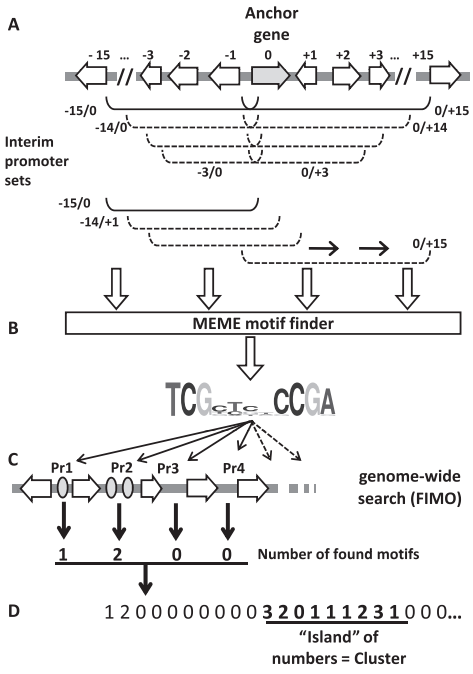

Fig. 2. CASSIS algorithm. (**A**) Interim promoter sets around the anchor gene are submitted to MEME for motif prediction. (**B**) All found motifs are selected. (**C**) The motifs are submitted to FIMO for the genome-wide prediction in promoter (Pr) sequences. (**D**) The sequence of promoters, each characterized by the number of found motifs, is considered as the string of numbers. This number string is searched for an 'island' of mostly non-zero values, which is regarded as the cluster

2.5.5 Adjustable parameters and their estimation

CASSIS can be fine tuned by adjusting four parameters, two of them being intrinsic CASSIS features, whereas the other two are the parameters of MEME and FIMO search. Since the motif prediction plays pivotal role in the further analysis, refining the latter by adjusting the *E*-value and *p*-value cut-offs for MEME and FIMO, respectively, can be crucial for the whole cluster prediction. The CASSIS default parameters for MEME and FIMO are estimated using the training set of experimentally verified SM clusters (Section 2.1). The option to tune these parameters is provided in CASSIS.

The two intrinsic CASSIS parameters are (i) the proportion of promoters with the motif in the genome (reflecting the genome-wide motif frequency); and (ii) the maximal allowed number of 'zero' promoters ('gaps') within the cluster. These parameters are estimated using a training set (e.g. of experimentally verified SM clusters) and can be further adjusted by the user. The gap parameter is restricted at the upper border by five promoters. The parameters are considered optimal if they give rise to the predictions with the smallest deviation. For the Ascomycete training set (Section 2.1), the parameter values were: frequency 14% and gap ≤2 zero-promoters (based on the observation of the largest gap in real clusters).

2.5.6 Runtime analysis

We applied CASSIS to the training set of 38 known gene clusters (Supplementary Table S1) on a machine with Intel Xeon CPUs running at 2.7 GHz. Using more than one CPU automatically turns on the parallelization of the MEME and FIMO steps. First, we allowed CASSIS to use up to 60 CPUs. Time measurements yield that it takes about 3 min in average to predict the cluster for a given anchor gene. Allowing only two CPUs, which should give results similar to a usual desktop computer, the prediction takes about 40 min in average.

3 Results

3.1 New features of CASSIS and prediction of SM enzymes by SMIPS

CASSIS is the improvement of the previously established MDM tool (Wolf *et al.*, 2013). CASSIS is not similarity based and exploits the properties used for the definition of clusters, namely the co-localization and presumable co-regulation of cluster genes. The co-regulation assumes the occurrence of binding sites for the common regulator (TF) in the promoters of cluster genes. The task of the cluster prediction is thus restricted to the task of finding a region around the anchor gene, where the promoters share a common binding site. Importantly, the promoters sharing the site should form an 'island'—a mostly uninterrupted group separated from non-cluster regions by long stretches of 'motif-less' promoters. This does not mean that the same sites cannot occur outside the cluster: they can exist but should be far enough not to interfere with the cluster (moreover, they can be indicators of other genes regulated by the same TF, thus the information about them can be valuable).

In the course of improvement of the MDM method and collecting more observations of real clusters, we realized that the

prediction algorithm can be simplified, so that the scoring system applied in the MDM version can be dropped. Instead of the 'frame scores' used in MDM, we now applied a more straightforward approach of 'gap rules' as described in Section 2.5.4. This made the algorithm more transparent, easier to adjust and easier to interpret. Besides this innovation, we added and improved several features. We drastically increased the number of promoter sets (from 7 to 250), which are submitted to MEME for motif prediction. This makes the search for the best common motif more precise, improving the accuracy of the entire cluster prediction. We also introduced several cut-off values to filter out unpromising or invalid intermediate results: (i) the *E*-value cut-off for motifs predicted by MEME, (ii) the *p*-value cut-off for FIMO hits, (iii) the percentage cut-off for the number of promoters with binding sites compared with all promoters in the genome (genome-wide frequency of the motif), and (iv) the maximal gap length within the cluster. Altogether this helped to decrease the number of FP and to increase the specificity and accuracy (see Section 3.2).

To make the workflow smooth and independent, we added a small but useful tool called SMIPS for the preliminary prediction of (all potential SM anchor genes in the given genome.) Methodologically, SMIPS does not differ from other tools for PKS and NRPS predictions, basing on the HMM models for typical domains of SM enzymes. However, as CASSIS requires predefined anchor genes as input, we found it more convenient to add SMIPS to the CASSIS workflow. In this way, we avoid preliminary runs of other tools (such as SMURF) to obtain the anchor genes information. In addition to sending the output of SMIPS to CASSIS, it can be used independently for the annotation and description of SM genes.

3.2 Assessment of the CASSIS performance, validation and comparison with other tools

To assess the performance of our method and tool, we undertook a series of leave-one-out (LOO) cross-validation experiments. As positive set we used the 38 experimentally proven fungal clusters (Supplementary Table S1) and performed the LOO for each cluster. The benchmarking shows high specificity, sensitivity, accuracy and precision (Table 1). With this we show that over-fitting is not an issue and our tool is able to reliably predict unknown clusters.

As the CASSIS approach is based on promoter analysis and is thus very distant from similarity-based methods, it was interesting to compare its performance with the most prominent similarity-based tools antiSMASH and SMURF. We applied the tools to the re-identification of the clusters with known borders. To make a clean experiment and put all three tools in equal position, we included in our training set those clusters that were characterized before the publication of antiSMASH and SMURF and hence could be used for their training (at least for SMURF). On the other hand, the clusters published after 2010 were considered as 'new' for all three tools and used as test set. The comparison reveals that antiSMASH has a higher sensitivity but the number of FP predictions made by similarity-based methods is also higher: compared with CASSIS, antiSMASH suggests in average four FP more per cluster. This reflects the tendency of the similarity tools to overestimate the clusters' lengths, even though they pick up the right genes with high sensitivity. As a result, CASSIS outperforms the other tools in specificity, accuracy and precision (Table 2, Supplementary Table S3 and Supplementary Fig. S1). Moreover, in some cases the similarity-based tools failed to recognize the anchor gene, which lead to the loss of the whole cluster (see Supplementary Fig. S2). For instance, in *Aspergillus nidulans* the *ent*-pimara-8(14),15-diene cluster is lost by antiSMASH and SMURF because they do not recognize AN1594 as the anchor gene. CASSIS/SMIPS did not encounter any problems in the detection of all anchors.

See Supplementary Table S4 for a more general comparison of the features of all four tools.

Table 1. Benchmark results of the LOO cross-validation for CASSIS

Characteristics	CASSIS performance[a]
Sensitivity	0.84 ± 0.0010
Specificity	0.98 ± 0.0002
Precision	0.71 ± 0.0010
Accuracy	0.96 ± 0.0002
FDR	0.29 ± 0.0010
F_1-score	0.73 ± 0.0008

[a]Average for all 38 LOO experiments. Error is the standard error of the mean. See Supplementary Table S1 for the list of used clusters

4 Discussion

Clustering of genes implies their co-localization, co-regulation and assignment to the same process. In the case of SM, this is a biosynthetic pathway and/or further processing of the product. Most of the approaches developed for the genome-wide cluster prediction rely on protein domain similarity. Thus, they use the first and last properties of the clusters but ignore the co-expression (or co-regulation). Our approach is in this sense complementary, as it ignores the functional features of the proteins but considers the promoter information. This constitutes both an advantage and a disadvantage of the approach. The advantage is the consideration of a new, yet unused layer of information (promoters, motifs, sites), which is, moreover, the key feature of the cluster definition. The disadvantage is the neglect of the remaining information, but this can be seen as a specialization. Indeed, the similarity-based tools exist and at least one of them, antiSMASH, gives very good, although not perfect, predictions. Our aim is not to compete with antiSMASH or to substitute it. We suppose that the optimal predictions can be achieved by application of both tools simultaneously: antiSMASH is more sensitive but CASSIS is more precise, and each of them supplies with specific information about the discovered clusters (see Supplementary Table S4).

Motifs that are shared by the cluster genes (and form in this way the basis of the cluster prediction) have their own value as the potential TFBS of the cluster's presumable regulator. CASSIS provides the option to retrieve the motifs corresponding to the detected clusters.

Table 2. Comparison of CASSIS with the similarity-based antiSMASH and SMURF tools: re-identification of the 12 test clusters not used for the tools' training

Characteristics	Comparison[a]		
	CASSIS	antiSMASH	SMURF
Sensitivity	0.87 ± 0.04	0.94 ± 0.04	0.78 ± 0.10
Specificity	0.96 ± 0.01	0.87 ± 0.02	0.84 ± 0.02
Precision	0.80 ± 0.05	0.54 ± 0.05	0.42 ± 0.06
Accuracy	0.94 ± 0.01	0.88 ± 0.01	0.82 ± 0.02
FDR	0.20 ± 0.05	0.46 ± 0.05	0.58 ± 0.06
F_1-score	0.81 ± 0.02	0.66 ± 0.04	0.51 ± 0.06

[a]Average for all 12 clusters. Error is the standard error of the mean. See Supplementary Table S1 for the list of used clusters

Moreover, as the motifs' occurrences are scanned genome-wide, it is possible to find sub-clusters (also called super-clusters), which are groups of genes regulated by the same TF, simultaneously with the 'main' cluster, but located in another part of the genome. If a sub-cluster is large enough (more than three genes), it can be detected quite easily. In the next versions of CASSIS we plan to implement such a feature.

Like MDM, CASSIS is not restricted to the prediction of SM clusters. Other types of gene clusters can be represented by different anchor genes, depending on the pathway or process, for which the genes are clustered. As CASSIS does not consider the properties of genes, the nature of the anchor gene does not matter.

Being based on the *de novo* motif discovery, CASSIS is quite sensitive to the quality of the genome assembly. Two features are important: the length of contigs (scaffolds) and the information quality of the sequence. The former feature is, actually, important for any cluster prediction tool, since clusters are lengthy stretches of genomic sequence, which should be preferably uninterrupted. The information quality becomes important for genomes with low complexity (AT-rich) regions, since it is hard to predict significant motifs in such sequences.

5 Implementation and availability

The CASSIS method is implemented in a tool with the same name. User-friendly online versions of both SMIPS and CASSIS (the 'CASSIS suite') are available at https://sbi.hki-jena.de/cassis. The suite also provides a comfortable workflow to run CASSIS on the results of SMIPS. The source codes as well as executeable files for Linux and Windows are freely available at https://sbi.hki-jena.de/cassis/Download.php. The SMIPS and CASSIS tools are implemented in Perl 5.

Acknowledgements

We would like to thank Zerrin Üzüm and Katharina Bonkowski for providing valuable suggestions on the SMIPS web page. Also, we thank Alina Burmistrova for technical assistance on the CASSIS and SMIPS web pages.

Funding

T.W. was supported by the International Leibniz Research School for Microbial and Molecular Interactions (ILRS), as part of the excellence graduate school Jena School for Microbial Communication (JSMC), supported by the Deutsche Forschungsgemeinschaft (DFG). This study was (in part) supported by the Collaborative Research Centre ChemBioSys (CRC 1127 ChemBioSys), funded by the DFG.

Conflict of interest: none declared.

References

Amaike,S. and Keller,N.P. (2011) *Aspergillus flavus*. In: VanAlfen N., *et al.* (ed.) *Annual Review of Phytopathology*. Vol. **49**, pp. 107–133.

Andersen,M.R. *et al.* (2013) Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc. Natl Acad. Sci.*, **110**, E99–E107.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings/International Conference on Intelligent Systems for Molecular Biology; ISMB*, Vol. 2, AAAI Press, pp. 28–36.

Bailey,T.L. *et al.* (2009) MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

Bergmann,S. *et al.* (2010) Activation of a silent fungal polyketide biosynthesis pathway through regulatory cross talk with a cryptic nonribosomal peptide synthetase gene cluster. *Appl. Environ. Microbiol.*, **76**, 8143–8149.

Blin,K. *et al.* (2013) antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.*, **41**, W204–W212.

Boettger,D. *et al.* (2012) Evolutionary imprint of catalytic domains in fungal PKS-NRPS hybrids. *ChemBioChem*, **13**, 2363–2373.

Brakhage,A.A. (2013) Regulation of fungal secondary metabolism. *Nat. Rev. Microbiol.*, **11**, 21–32.

Brakhage,A.A. and Schroeckh,V. (2011) Fungal secondary metabolites—strategies to activate silent gene clusters. *Fungal Genet. Biol.*, **48**, 15–22.

Cerqueira,G.C. *et al.* (2014) The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Res.*, **42**, D705–D710.

Do,J. and Miyano,S. (2008) The GC and window-averaged DNA curvature profile of secondary metabolite gene cluster in *Aspergillus fumigatus*y genome. *Appl. Microbiol. Biotechnol.*, **80**, 841–847.

Eppelmann,K. *et al.* (2002) Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry*, **41**, 9718–9726.

Grant,C.E. *et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Hoffmeister,D. and Keller,N.P. (2007) Natural products of filamentous fungi: enzymes, genes, and their regulation. *Nat. Prod. Rep.*, **24**, 393–416.

Keller,N.P. and Hohn,T.M. (1997) Metabolic pathway gene clusters in filamentous fungi. *Fungal Genet. Biol.: FG & B*, **21**, 17–29.

Khaldi,N. *et al.* (2010) SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.*, **47**, 736–741.

Stachelhaus,T. *et al.* (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, **6**, 493–505.

Starcevic,A. *et al.* (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. *Nucleic Acids Res.*, **36**, 6882–6892.

Weber,T. *et al.* (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.*, **140**, 13–17.

Wolf,T. *et al.* (2013) Motif-based method for the genome-wide prediction of eukaryotic gene clusters. In: Petrosino A., *et al.* (eds), *New Trends in Image Analysis and Processing – ICIAP 2013, Number 8158 in Lecture Notes in Computer Science*, Springer, Heidelberg, pp. 389–398.

2.4.1 Ergänzende Materialien zu Manuskript 3

Table S1: Set of 38 secondary metabolites, their anchor genes, and known (experimentally proven) cluster borders. For comparing CASSIS with antiSMASH and SMURF, blue clusters are used for parameter estimation/training and red clusters for validation/testing.

Species	Secondary metabolite / gene cluster	Anchor gene	Known cluster borders first gene	last gene	Ref.
A. clavatus	Cytochalasin	Acla078660	Acla078640	Acla078710	[21]
A. flavus	Aflatoxin[a]	Afl2g07228	Afl2g07209	Afl2g07230	[33]
A. fumigatus	Conidial pigment	Afu2g17600	Afu2g17530	Afu2g17600	[27]
A. fumigatus	Fumagillin[b]	Afu8g00370	Afu8g00370	Afu8g00520	[14]
A. fumigatus	Fumicycline[c]	Afu7g00160	Afu7g00120	Afu7g00180	[10]
A. fumigatus	Fumigaclavine	Afu2g18040	Afu2g17960	Afu2g18060	[29]
A. fumigatus	Fumiquinazoline	Afu6g12050	Afu6g12040	Afu6g12110	[1]
A. fumigatus	Fumitremorgin	Afu8g00170	Afu8g00170	Afu8g00250	[16]
A. fumigatus	Gliotoxin	Afu6g09660	Afu6g09630	Afu6g09745	[23]
A. fumigatus	Hexadehydroastechrome	Afu3g12920	Afu3g12890	Afu3g12960	[32]
A. fumigatus	Pseurotin	Afu8g00540	Afu8g00540	Afu8g00570	[28]
A. graminea.	Aurofusarin[a]	Fgsg02324	Fgsg02320	Fgsg02330	[17]
A. nidulans	Asperfuranone	An1034	An1036	An1029	[8]
A. nidulans	Asperfuranone	An1036	An1036	An1029	[8]
A. nidulans	Aspernidine	An3230	An3230	An3225	[30]
A. nidulans	Asperthecin	An6000	An6002	An6000	[25]
A. nidulans	Aspyridone	An8412	An8408	An8415	[3]
A. nidulans	Austinol A	An8383	An8379	An8384	[15]
A. nidulans	Austinol B	An9259	An9246	An9259	[15]
A. nidulans	Cichorine	An6448	An6449	An6443	[22]
A. nidulans	Emericellamide	An2545	An2545	An2549	[7]
A. nidulans	Emericellamide	An2547	An2545	An2549	[7]
A. nidulans	*ent*-pimara[d]	An1594	An1592	An1599	[6]
A. nidulans	inp	An3496	An3496	An3490	[4]
A. nidulans	Monodictyphenone	An0150	An10023	An10021	[9]
A. nidulans	Orsellinic acid	An7909	An7909	An7914	[24]
A. nidulans	Penicillin	An2621	An2623	An2621	[18]
A. nidulans	Sterigmatocystin	An7825	An7804	An7825	[33]
A. nidulans	Terriquinone	An8514	An8513	An8520	[5]
A. nidulans	Violaceol	An7903	An7896	An7903	[11]
A. niger	Pyranonigrin	An11g00250	An11g00250	An11g00350	[2]
A. oryzae	Aflatoxin	Ao090026000009	Ao090026000032	Ao090026000008	[26]
A. oryzae	AOI	Ao090010000048	Ao090010000056	Ao090010000040	[19]
A. oryzae	WYK	Ao090001000009	Ao090001000009	Ao090001000019	[13]
A. terreus	Acetylaranotin	Ateg03470	Ateg03466	Ateg03475	[12]
A. terreus	Geodin	Ateg08451	Ateg08449	Ateg08460	[20]
A. terreus	Terrein	Ateg00145	Ateg00135	Ateg00145	[34]
N. fischeri	Acetylaszonalenin[b]	Nfia055290	Nfia055290	Nfia055310	[31]

[a] no suitable input file format available for antiSMASH
[b] cluster not found by antiSMASH
[c] Fumicycline from *Aspergillus fumigatus*, Neosartoricin from *Neosartorya fischeri*
[d] *ent*-pimara-8(14),15-diene; anchor gene not found by antiSMASH and SMURF, but SMIPS

Table S2: Mapping of InterPro IDs (IPR) to protein domains and their corresponding anchor gene types.

InterPro ID (IPR)	Protein domain type[a]	Anchor gene type (typically)
IPR014043	AT	PKS
IPR016035	AT	
IPR016036	AT	
IPR020801	AT	
IPR006765	Cyc	
IPR020807	DH	
IPR011032	ER	
IPR013149	ER	
IPR013154	ER	
IPR020843	ER	
IPR013968	KR	
IPR020842	KR	
IPR014030	KS	
IPR014031	KS	
IPR018201	KS	
IPR020841	KS	
IPR004033	MT	
IPR013216	MT	
IPR013217	MT	
IPR020803	MT	
IPR029063	MT	
IPR015083	xPKS	
IPR013601	xPKS	
IPR000873	A	NRPS
IPR010071	A	
IPR020845	A	
IPR025110	A	
IPR001242	C	
IPR010060	xNRPS	
IPR012728	xNRPS	
IPR013624	xNRPS	
IPR003231	ACP	PKS or NRPS
IPR004568	ACP	
IPR006162	ACP	
IPR008278	ACP	
IPR009081	ACP	
IPR020806	ACP	
IPR001031	TE	
IPR010080	TE	
IPR020802	TE	
IPR000537	Prenyltransferase	DMATS
IPR012148	Prenyltransferase	
IPR017795	Prenyltransferase	
IPR017796	Prenyltransferase	

[a] AT: Acyl transferase; ACP ($\hat{=}$ PCP $\hat{=}$ PP): Acyl carrier protein, Peptidyl carrier protein, Phosphopantetheine; KS: Beta-ketoacyl synthase; KR: Keto reductase; DH: Dehydratase; ER: Enoylreductase, GroES-like, Alcohol dehydrogenase; MT: Methyltransferase; TE: Thioesterase, Thioester reductase-like; CYC: Cyclase; xPKS: Typical PKS domain with yet unkown function; A: Adenylation, AMP-dependent synthetase/ligase; C: Condensation; xNRPS: Typical NRPS domain with yet unkown function

Table S3: Statistical characteristics comparison of cluster re-predictions by CASSIS, antiSMASH, and SMURF.

Characteristics[a,b]	Sens.	Spec.	Prec.	FDR	Acc.	F_1	$F_{0.5}$	F_2
Entire training set, 33[c] clusters								
CASSIS	0.84	0.96	0.72	0.28	0.94	0.74	0.72	0.79
antiSMASH	0.94	0.85	0.46	0.54	0.85	0.59	0.50	0.73
SMURF	0.89	0.88	0.53	0.47	0.87	0.62	0.56	0.73
Leave-one-out, 33[c] clusters, no prediction for inp cluster by CASSIS								
CASSIS	0.82	0.96	0.72	0.28	0.94	0.72	0.71	0.77
antiSMASH	0.94	0.85	0.46	0.54	0.85	0.59	0.50	0.73
SMURF	0.89	0.88	0.53	0.47	0.87	0.62	0.56	0.73
Published 2011 or later, excluding *Aspergillus fumigatus*, 12[c] clusters								
CASSIS	0.87	0.96	0.80	0.20	0.94	0.81	0.80	0.84
antiSMASH	0.94	0.87	0.54	0.46	0.88	0.66	0.58	0.79
SMURF	0.78	0.84	0.42	0.58	0.82	0.51	0.44	0.63

[a] Sens: sensitivity; Spec: specificity; Prec: precision; FDR: false discovery rate; Acc: accuracy; F_1-score: weighted average of precision and sensitivity; $F_{0.5}$-score: preferring precision; F_2-score: preferring sensitivity

[b] average for all clusters

[c] some clusters had to be discarded from the statistical measurement, hence sets contain only 33 instead of 38 and 12 instead of 14 clusters; see Table S1

Table S4: Feature comparison of CASSIS (SMIPS), antiSMASH, and SMURF.

	CASSIS	antiSMASH	SMURF
Based on	transcription factor binding sites (SMIPS: protein domains)	protein domains	protein domains
Applicable to (genomes)			
prokaryotic	no (SMIPS: yes)	yes	no
eukaryotic	yes	yes	yes (fungi only)
Applicable to (anchor genes)			
NRPS	yes	yes	yes
PKS	yes	yes	yes
NRPS-PKS hybrids	yes	yes	yes
DMATS	yes	yes	yes
others	yes (SMIPS: no)	yes	no
Input type[a]	CASSIS: • genome • annotation • anchor gene SMIPS: • InterProScan results or proteins	• genome or proteins	• proteins • annotation
multi-contig input	yes	yes	yes
Results provided			
anchor genes	yes (by SMIPS)	yes	yes
domain arrangement	yes (by SMIPS)	sometimes	no
domain annotation	yes (by SMIPS)	yes	yes
product's core structure	no	yes	no
cluster borders	yes	yes	yes
comparative cluster analysis	no	yes	no
sub-cluster analysis	no	yes	no
binding site motifs	yes	no	no
promoter sequences	yes	no	no
Availability			
web-service	yes (CASSIS: in preparation)	yes	yes
source code	yes	yes	no
binaries	yes	yes	no

[a] genome: nucleotide sequences; proteins: amino acid sequences; annotation: feature annotation

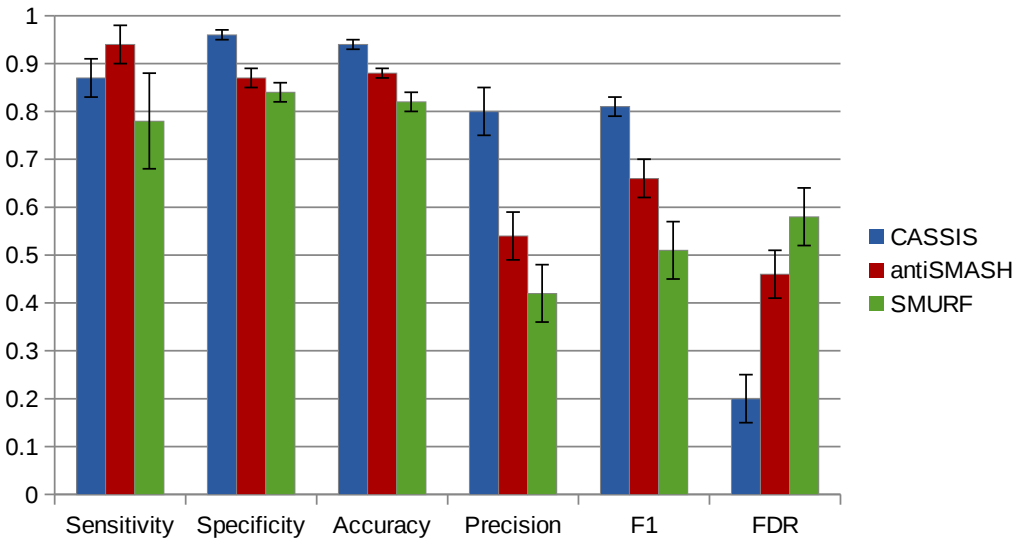

Figure S1: Comparison of CASSIS with the similarity-based antiSMASH and SMURF tools: Re-identification of the 12 test clusters not used for the tools' training (published 2011 or later, excluding *Aspergillus fumigatus*). Error bars show the standard error of the mean.

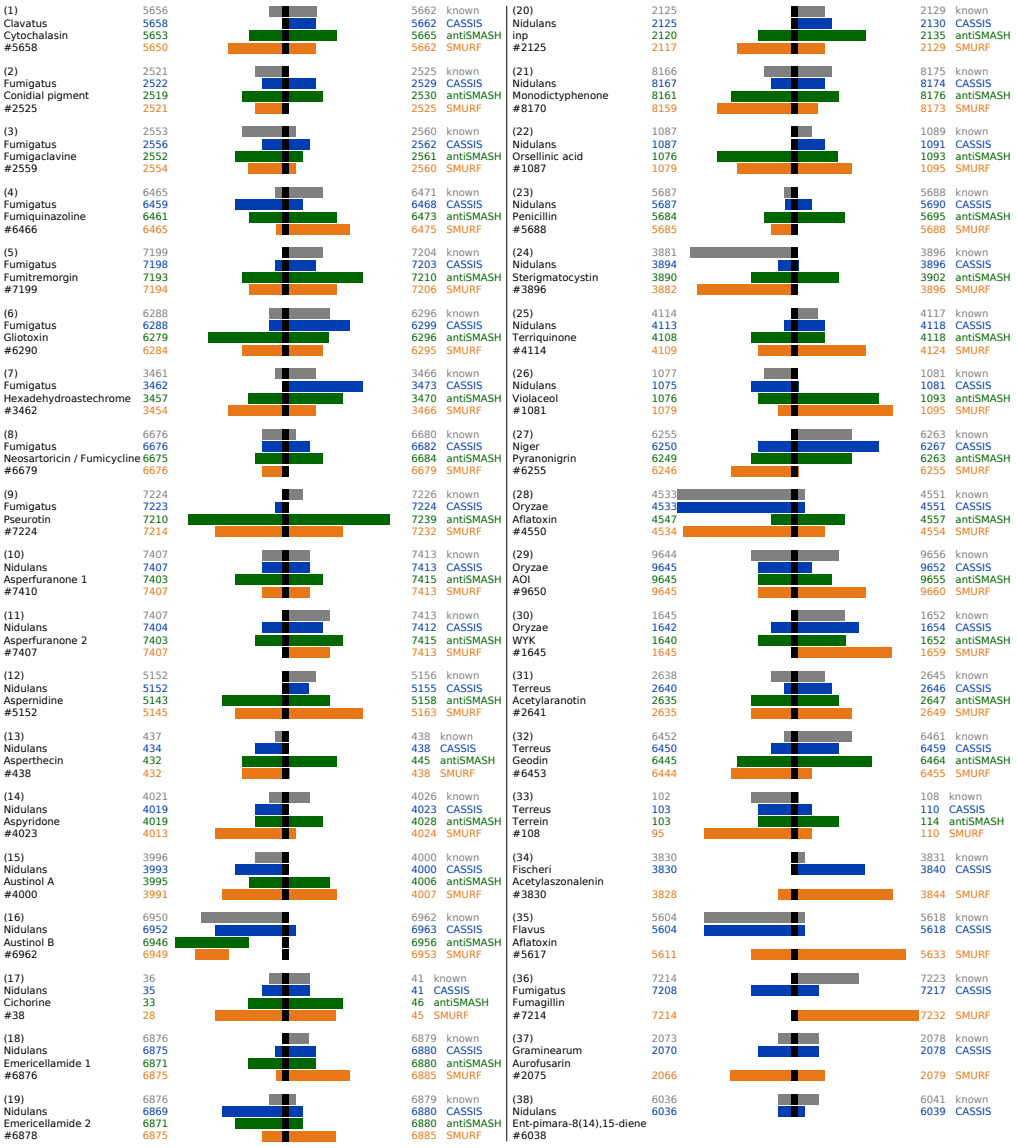

Figure S2: Comparison of CASSIS with the similarity-based antiSMASH and SMURF tools: Re-identification of 38 known clusters. The promoter numbers of the anchor genes (#) and the upstream/downstream cluster borders are given.

References

[1] Ames, B. D., Liu, X., and Walsh, C. T. (2010). Enzymatic Processing of Fumiquinazoline F: A Tandem Oxidative-Acylation Strategy for the Generation of Multicyclic Scaffolds in Fungal Indole Alkaloid Biosynthesis. *Biochemistry*, **49**(39), 8564–8576.

[2] Awakawa, T., Yang, X.-L., Wakimoto, T., and Abe, I. (2013). Pyranonigrin E: A PKS-NRPS Hybrid Metabolite from *Aspergillus niger* Identified by Genome Mining. *ChemBioChem*, **14**(16), 2095–2099.

[3] Bergmann, S., Schumann, J., Scherlach, K., Lange, C., Brakhage, A. A., and Hertweck, C. (2007). Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. *Nat Chem Biol*, **3**(4), 213–217.

[4] Bergmann, S., Funk, A. N., Scherlach, K., Schroeckh, V., Shelest, E., Horn, U., Hertweck, C., and Brakhage, A. A. (2010). Activation of a Silent Fungal Polyketide Biosynthesis Pathway through Regulatory Cross Talk with a Cryptic Nonribosomal Peptide Synthetase Gene Cluster. *Applied and Environmental Microbiology*, **76**(24), 8143–8149.

[5] Bouhired, S., Weber, M., Kempf-Sontag, A., Keller, N. P., and Hoffmeister, D. (2007). Accurate prediction of the *Aspergillus nidulans* terrequinone gene cluster boundaries using the transcriptional regulator LaeA. *Fungal Genetics and Biology*, **44**(11), 1134–1145.

[6] Bromann, K., Toivari, M., Viljanen, K., Vuoristo, A., Ruohonen, L., and Nakari-Setälä, T. (2012). Identification and Characterization of a Novel Diterpene Gene Cluster in *Aspergillus nidulans*. *PLoS ONE*, **7**(4), e35450.

[7] Chiang, Y.-M., Szewczyk, E., Nayak, T., Davidson, A. D., Sanchez, J. F., Lo, H.-C., Ho, W.-Y., Simityan, H., Kuo, E., Praseuth, A., Watanabe, K., Oakley, B. R., and Wang, C. C. (2008). Molecular Genetic Mining of the Aspergillus Secondary Metabolome: Discovery of the Emericellamide Biosynthetic Pathway. *Chemistry & Biology*, **15**(6), 527–532.

[8] Chiang, Y.-M., Szewczyk, E., Davidson, A. D., Keller, N., Oakley, B. R., and Wang, C. C. C. (2009). A Gene Cluster Containing Two Fungal Polyketide Synthases Encodes the Biosynthetic Pathway for a Polyketide, Asperfuranone, in *Aspergillus nidulans*. *J. Am. Chem. Soc.*, **131**(8), 2965–2970.

[9] Chiang, Y.-M., Szewczyk, E., Davidson, A. D., Entwistle, R., Keller, N. P., Wang, C. C. C., and Oakley, B. R. (2010). Characterization of the *Aspergillus nidulans* Monodictyphenone Gene Cluster. *Applied and Environmental Microbiology*, **76**(7), 2067–2074.

[10] Chooi, Y.-H., Fang, J., Liu, H., Filler, S. G., Wang, P., and Tang, Y. (2013). Genome Mining of a Prenylated and Immunosuppressive Polyketide from Pathogenic Fungi. *Organic Letters*, **15**(4), 780–783.

[11] Gerke, J., Bayram, , Feussner, K., Landesfeind, M., Shelest, E., Feussner, I., and Braus, G. H. (2012). Breaking the Silence: Protein Stabilization Uncovers Silenced Biosynthetic Gene Clusters in the Fungus *Aspergillus nidulans*. *Applied and Environmental Microbiology*, **78**(23), 8234–8244.

[12] Guo, C.-J., Yeh, H.-H., Chiang, Y.-M., Sanchez, J. F., Chang, S.-L., Bruno, K. S., and Wang, C. C. C. (2013). Biosynthetic Pathway for the Epipolythiodioxopiperazine Acetylaranotin in *Aspergillus terreus* Revealed by Genome-Based Deletion Analysis. *Journal of the American Chemical Society*, **135**(19), 7205–7213.

[13] Imamura, K., Tsuyama, Y., Hirata, T., Shiraishi, S., Sakamoto, K., Yamada, O., Akita, O., and Shimoi, H. (2012). Identification of a Gene Involved in the Synthesis of a Dipeptidyl Peptidase IV Inhibitor in *Aspergillus oryzae*. *Applied and Environmental Microbiology*, **78**(19), 6996–7002.

[14] Lin, H.-C., Chooi, Y.-H., Dhingra, S., Xu, W., Calvo, A. M., and Tang, Y. (2013). The Fumagillin Biosynthetic Gene Cluster in *Aspergillus fumigatus* Encodes a Cryptic Terpene Cyclase Involved in the Formation of β-trans-Bergamotene. *Journal of the American Chemical Society*, **135**(12), 4616–4619.

[15] Lo, H.-C., Entwistle, R., Guo, C.-J., Ahuja, M., Szewczyk, E., Hung, J.-H., Chiang, Y.-M., Oakley, B. R., and Wang, C. C. C. (2012). Two Separate Gene Clusters Encode the Biosynthetic Pathway for the Meroterpenoids Austinol and Dehydroaustinol in *Aspergillus nidulans*. *Journal of the American Chemical Society*, **134**(10), 4709–4720.

[16] Maiya, S., Grundmann, A., Li, S.-M., and Turner, G. (2006). The Fumitremorgin Gene Cluster of *Aspergillus fumigatus*: Identification of a Gene Encoding Brevianamide F Synthetase. *ChemBioChem*, **7**(7), 1062–1069.

[17] Malz, S., Grell, M. N., Thrane, C., Maier, F. J., Rosager, P., Felk, A., Albertsen, K. S., Salomon, S., Bohn, L., Schäfer, W., and Giese, H. (2005). Identification of a gene cluster responsible for the biosynthesis of aurofusarin in the *Fusarium graminearum* species complex. *Fungal Genetics and Biology*, **42**(5), 420–433.

[18] Martín, J. F. (2000). Molecular Control of Expression of Penicillin Biosynthesis Genes in Fungi: Regulatory Proteins Interact with a Bidirectional Promoter Region. *Journal of Bacteriology*, **182**(9), 2355–2362.

[19] Nakazawa, T., Ishiuchi, K., Praseuth, A., Noguchi, H., Hotta, K., and Watanabe, K. (2012). Overexpressing Transcriptional Regulator in *Aspergillus oryzae* Activates a Silent Biosynthetic Pathway to Produce a Novel Polyketide. *ChemBioChem*, **13**(6), 855–861.

[20] Nielsen, M. T., Nielsen, J. B., Anyaogu, D. C., Holm, D. K., Nielsen, K. F., Larsen, T. O., and Mortensen, U. H. (2013). Heterologous Reconstitution of the Intact Geodin Gene Cluster in *Aspergillus nidulans* through a Simple and Versatile PCR Based Approach. *PLoS ONE*, **8**(8), e72871.

[21] Qiao, K., Chooi, Y.-H., and Tang, Y. (2011). Identification and engineering of the cytochalasin gene cluster from *Aspergillus clavatus* NRRL 1. *Metabolic Engineering*, **13**(6), 723–732.

[22] Sanchez, J. F., Entwistle, R., Corcoran, D., Oakley, B. R., and Wang, C. C. C. (2012). Identification and molecular genetic analysis of the cichorine gene cluster in *Aspergillus nidulans*. *MedChemComm*, **3**(8).

[23] Scharf, D. H., Heinekamp, T., Remme, N., Hortschansky, P., Brakhage, A. A., and Hertweck, C. (2011). Biosynthesis and function of gliotoxin in *Aspergillus fumigatus*. *Applied Microbiology and Biotechnology*, **93**(2), 467–472.

[24] Schroeckh, V., Scherlach, K., Nützmann, H.-W., Shelest, E., Schmidt-Heck, W., Schuemann, J., Martin, K., Hertweck, C., and Brakhage, A. A. (2009). Intimate bacterial–fungal interaction triggers biosynthesis of archetypal polyketides in *Aspergillus nidulans*. *Proceedings of the National Academy of Sciences*, **106**(34), 14558–14563.

[25] Szewczyk, E., Chiang, Y.-M., Oakley, C. E., Davidson, A. D., Wang, C. C. C., and Oakley, B. R. (2008). Identification and Characterization of the Asperthecin Gene Cluster of *Aspergillus nidulans*. *Applied and Environmental Microbiology*, **74**(24), 7607–7612.

[26] Tominaga, M., Lee, Y.-H., Hayashi, R., Suzuki, Y., Yamada, O., Sakamoto, K., Gotoh, K., and Akita, O. (2006). Molecular Analysis of an Inactive Aflatoxin Biosynthesis Gene Cluster in *Aspergillus oryzae* RIB Strains. *Applied and Environmental Microbiology*, **72**(1), 484 –490.

[27] Tsai, H.-F., Wheeler, M. H., Chang, Y. C., and Kwon-Chung, K. J. (1999). A Developmentally Regulated Gene Cluster Involved in Conidial Pigment Biosynthesis in *Aspergillus fumigatus*. *Journal of Bacteriology*, **181**(20), 6469–6477.

[28] Vödisch, M., Scherlach, K., Winkler, R., Hertweck, C., Braun, H.-P., Roth, M., Haas, H., Werner, E. R., Brakhage, A. A., and Kniemeyer, O. (2011). Analysis of the *Aspergillus fumigatus* Proteome Reveals Metabolic Changes and the Activation of the Pseurotin A Biosynthesis Gene Cluster in Response to Hypoxia. *J. Proteome Res.*, **10**(5), 2508–2524.

[29] Wallwey, C., Matuschek, M., Xie, X.-L., and Li, S.-M. (2010). Ergot alkaloid biosynthesis in *Aspergillus fumigatus*: Conversion of chanoclavine-I aldehyde to festuclavine by the festuclavine synthase FgaFS in the presence of the old yellow enzyme FgaOx3. *Organic & Biomolecular Chemistry*, **8**(15), 3500–3508.

[30] Yaegashi, J., Praseuth, M. B., Tyan, S.-W., Sanchez, J. F., Entwistle, R., Chiang, Y.-M., Oakley, B. R., and Wang, C. C. C. (2013). Molecular Genetic Characterization of the Biosynthesis Cluster of a Prenylated Isoindolinone Alkaloid Aspernidine A in *Aspergillus nidulans*. *Organic Letters*, **15**(11), 2862–2865.

[31] Yin, W.-B., Grundmann, A., Cheng, J., and Li, S.-M. (2009). Acetylaszonalenin Biosynthesis in *Neosartorya fischeri*. Identification of the biosynthetic gene cluster by genomic mining and functional proof of the genes by biochemical investigation. *Journal of Biological Chemistry*, **284**(1), 100–109.

[32] Yin, W.-B., Baccile, J. A., Bok, J. W., Chen, Y., Keller, N. P., and Schroeder, F. C. (2013). A Nonribosomal Peptide Synthetase-Derived Iron(III) Complex from the Pathogenic Fungus *Aspergillus fumigatus*. *Journal of the American Chemical Society*, **135**(6), 2064–2067.

[33] Yu, J., Chang, P.-K., Ehrlich, K. C., Cary, J. W., Bhatnagar, D., Cleveland, T. E., Payne, G. A., Linz, J. E., Woloshuk, C. P., and Bennett, J. W. (2004). Clustered Pathway Genes in Aflatoxin Biosynthesis. *Applied and Environmental Microbiology*, **70**(3), 1253 –1262.

[34] Zaehle, C., Gressler, M., Shelest, E., Geib, E., Hertweck, C., and Brock, M. (2014). Terrein Biosynthesis in *Aspergillus terreus* and Its Impact on Phytotoxicity. *Chemistry & Biology*, **21**(6), 719–731.

2.5 Manuskript 4: »Genome Sequences of Three *Pseudoalteromonas* Strains (P1-8, P1-11, and P1-30), Isolated from the Marine Hydroid *Hydractinia echinata*«

Status veröffentlicht, Dezember 2015

Literaturangabe Klassen, Jonathan L. ; Rischer, Maja ; Wolf, Thomas ; Guo, Huijuan ; Shelest, Ekaterina ; Clardy, Jon ; Beemelmanns, Christine: Genome Sequences of Three *Pseudoalteromonas* Strains (P1-8, P1-11, and P1-30), Isolated from the Marine Hydroid *Hydractinia echinata*. In: *Genome Announcements* 3 (2015), Dezember, Nr. 6, e01380–15. `http://dx.doi.org/10.1128/genomeA.01380-15`. – DOI 10.1128/genomeA.01380–15. – ISSN 2169–8287

Übersicht *Pseudoalteromonas* ist ein marines Bakterium, das meist im Verbund mit Eukaryoten lebt. Es ist bekannt für die Biosynthese von antimikrobiellen Wirkstoffen. Es wurden drei *Pseudoalteromonas*-Stämme aus dem Gewebe der marinen Hydrozoe *Hydractinia echinata* isoliert und deren Genome sequenziert. Dabei wurden Gene für die Biofilmbildung, für die Anhaftung an Oberflächen und für die Synthese von Sekundärmetaboliten (genomweite Suche nach PKSs, NRPSs und DMATSs) identifiziert. Mit Hilfe der Sequenzdaten können diese Mechanismen besser verstanden und neue pharmazeutische Wirkstoffe entwickelt werden.

Beiträge JC und CB konzipierten, planten und organisierten die Experimente. MR, HJ und CB führten die Experimente durch. TW und JK waren für die Datenanalyse zuständig. TW analysierte die Sequenzdaten in Hinblick auf die Biosynthese von Sekundärmetaboliten und die Genomkomposition. JC, ES und CB stellten Material und Analysewerkzeuge zur Verfügung. MR, TW und CB schrieben die Publikation.

genomeA
Journals.ASM.org

Draft Genome Sequences of Six *Pseudoalteromonas* Strains, P1-7a, P1-9, P1-13-1a, P1-16-1b, P1-25, and P1-26, Which Induce Larval Settlement and Metamorphosis in *Hydractinia echinata*

Jonathan L. Klassen,[a] Thomas Wolf,[b] Maja Rischer,[b] Huijuan Guo,[b] Ekaterina Shelest,[b] Jon Clardy,[c] Christine Beemelmanns[b]

University of Connecticut, Department of Molecular & Cell Biology, Storrs, Connecticut, USA[a]; Leibniz Institute for Natural Product Research and Infection Biology eV, Jena, Germany[b]; Harvard Medical School, Department of Biological Chemistry and Molecular Pharmacology, Boston, Massachusetts, USA[c]

To gain a broader understanding of the importance of a surface-associated lifestyle and morphogenic capability, we have assembled and annotated the genome sequences of *Pseudoalteromonas* strains P1-7a, P1-9, P1-13-1a, P1-16-1b, P1-25, and P1-26, isolated from *Hydractinia echinata*. These genomes will allow detailed studies on bacterial factors mediating interkingdom communication.

Received 27 October 2015 **Accepted** 30 October 2015 **Published** 17 December 2015

Citation Klassen JL, Wolf T, Rischer M, Guo H, Shelest E, Clardy J, Beemelmanns C. 2015. Draft genome sequences of six *Pseudoalteromonas* strains, P1-7a, P1-9, P1-13-1a, P1-16-1b, P1-25, and P1-26, which induce larval settlement and metamorphosis in *Hydractinia echinata*. Genome Announc 3(6):e01477-15. doi:10.1128/genomeA.01477-15.

Address correspondence to Christine Beemelmanns, christine.beemelmanns@hki-jena.de.

Pseudoalteromonas strains P1-7a, P1-9, P1-13-1a, P1-16-1b, P1-25, and P1-26 were isolated from the tissue of a feeding polyp of the marine hydroid *Hydractinia echinata* (1) purchased from the Marine Biological Laboratory in Woods Hole, MA, USA. Pseudoalteromonads are commonly isolated from biofilms of marine surfaces and host tissue of marine invertebrates (2, 3). Their effects on the settlement and metamorphosis of biofouling invertebrates (4–6) and the production of pharmacologically active compounds (7) have been extensively studied. Six *Pseudoalteromonas* strains were isolated from *H. echinata* and screened for their effects on its larval settlement and metamorphosis using a colony-based assay (1). Genomes from the most inductive strains P1-7a, P1-9, P1-13-1a, P1-16-1b, P1-25, and P1-26 were sequenced to identify candidate genes responsible for larval settlement. Genomic DNA was extracted using the GenElute Blood Genomic DNA kit (Sigma-Aldrich) according to the manufacturer's protocol. Sequencing performed at the Harvard Medical School Biopolymers Facility used Illumina TruSeq 50 bp single-read libraries and a HiSeq2000 instrument (Illumina CASAVA 1.8.2). After subsampling reads to achieve ~50× coverage, genomes were assembled using the A5 pipeline v20120518 (8) and screened for contamination using blobology (9). Genomes were annotated using Prokka v1.10 (10) and assembly statistics were calculated using scripts from the Assemblathon2 project (11).

The draft genome sequence of strain P1-7a was sequenced to 52× coverage, and comprises 189 contigs totaling 4,374,565 bases in length and having a G+C content of 40.8%. Its annotation includes 3,853 coding sequences (CDSs), 96 tRNAs, and 4 rRNAs.

The draft genome of strain P1-9 was sequenced to 47× coverage, and comprises 211 contigs totaling 4,808,111 bases in length and having a G+C content of 40.7%. Its annotation includes 4,321 CDSs, 84 tRNAs, and 3 rRNAs.

The draft genome sequence of strain P1-13-1a was sequenced to 51× coverage, and comprises 174 contigs totaling 4,442,776 bases in length and having a G+C content of 40.7%. Its annotation includes 3,930 CDSs, 93 tRNAs, and 3 rRNAs.

The draft genome of strain P1-16-1b was sequenced to 57× coverage, and comprises 90 contigs totaling 3,977,637 bases in length and having a G+C content of 40.1%. Its annotation includes 3,562 CDSs, 90 tRNAs, and 4 rRNAs.

The draft genome sequence of strain P1-25 was sequenced to 51× coverage, and comprises 163 contigs totaling 4,399,610 bases in length and having a G+C content of 40.7%. Its annotation includes 3,855 CDS, 97 tRNAs, and 3 rRNAs.

The draft genome sequence of strain P1-26 was sequenced to 48× coverage, and comprises 219 contigs totaling 4,715,935 bases in length and having a G+C content of 41.2%. Its annotation includes 4,183 CDS, 96 tRNAs, and 4 rRNAs.

Genes associated with secretion (e.g., type II secretion system), biofilm formation (e.g., curli, extracellular polymers) (12), secondary metabolite production (e.g., NRPS), siderophore (e.g., desferrioxamine) (13, 14), and bacteriocin biosynthesis were detected in all genomes indicating the successful adaptation to persistence and competition on marine surfaces. These genome sequences will help elucidate the mechanisms involved in *H. echinata* settlement and metamorphosis (1), and help identify novel biotechnologically important molecules.

Nucleotide sequence accession numbers. These whole-genome shotgun projects for strains P1-7a, P1-9, P1-13-1a, P1-16-1b, P1-25, and P1-26 have been deposited in DDBJ/EMBL/GenBank under the accession numbers LKDU00000000, LKBD00000000, LKDV00000000, LKGQ00000000, LKDW00000000, and LKDX00000000, respectively. The versions described in this paper are the first versions, LKDU01000000, LKBD01000000, LKDV01000000, LKGQ01000000, LKDW01000000, and LKDX01000000.

ACKNOWLEDGMENTS

We are grateful for financial support from the NIH to J.C. (GM086258), and the German National Academy of Sciences Leopoldina for a postdoctoral fellowship to C.B. (LPDS 2011-2). J.L.K. was supported by funds from the University of Connecticut. M.R. was supported by the graduate school Jena School for Microbial Communication (JSMC) financed by the Deutsche Forschungsgemeinschaft, and T.W. was supported by the International Leibniz Research School for Microbial and Molecular Interactions (ILRS), as part of the JSMC.

REFERENCES

1. **Frank U, Leitz T, Müller WA**. 2001. The hydroid *Hydractinia*: a versatile, informative cnidarian representative. Bioessays **23:**963–971. http://dx.doi.org/10.1002/bies.1137.
2. **Gauthier G, Gauthier M, Christen R**. 1995. Phylogenetic analysis of the genera *Alteromonas*, *Shewanella*, and *Moritella* using genes coding for small-subunit rRNA sequences and division of the genus *Alteromonas* into two genera, *Alteromonas* (emended) and *Pseudoalteromonas* gen. nov., and proposal of twelve new species combinations. Int J Syst Bacteriol **45:**755–761. http://dx.doi.org/10.1099/00207713-45-4-755.
3. **Holmstrom C, Kjelleberg S**. 1999. Marine *Pseudoalteromonas* species are associated with higher organisms and produce biologically active extracellular agents. FEMS Microbiol Ecol **30:**285–293. http://dx.doi.org/10.1016/S0168-6496(99)00063-X
4. **Bowman JP**. 2007. Bioactive compound synthetic capacity and ecological significance of marine bacterial genus *Pseudoalteromonas*. Mar Drugs **5:**220–241. http://dx.doi.org/10.3390/md504220.
5. **Qian P-, Lau SCK, Dahms H-, Dobretsov S, Harder T**. 2007. Marine biofilms as mediators of colonization by marine macroorganisms: implications for antifouling and aquaculture. Mar Biotechnol **9:**399–410. http://dx.doi.org/10.1007/s10126-007-9001-9.
6. **Hadfield MG**. 2011. Biofilms and marine invertebrate larvae: what bacteria produce that larvae use to choose settlement sites. Annu Rev Marine Sci **3:**453–470. http://dx.doi.org/10.1146/annurev-marine-120709-142753.
7. **Machado H, Sonnenschein EC, Melchiorsen J, Gram L**. 2015. Genome mining reveals unlocked bioactive potential of marine gram-negative bacteria. BMC Genomics **16:**158–170. http://dx.doi.org/10.1186/s12864-015-1365-z.
8. **Tritt A, Eisen JA, Facciotti MT, Darling AE**. 2012. An integrated pipeline for *de novo* assembly of microbial genomes. PLoS One **7:**e42304. http://dx.doi.org/10.1371/journal.pone.0042304.
9. **Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M**. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet **4:**237. http://dx.doi.org/10.3389/fgene.2013.00237.
10. **Seemann T**. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30:**2068–2069. http://dx.doi.org/10.1093/bioinformatics/btu153.
11. **Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou W, Corbeil J, Del Fabbro C, Docking T, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang H, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam T-W, Lavenier D, Laviolette F, Li Y, Li Z, Lio B, Liu Y, Luo R, MacCallum I, MacManes MD, Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu S-M, Yuan J, Zhang G, Zhang H, Zhou S, Korf IF**. 2013. Assemblathon 2: Evaluating *de novo* methods of genome assembly in three vertebrate species. GigaScience **2:**10. http://dx.doi.org/10.1186/2047-217X-2-10.
12. **Thomas T, Evans FF, Schleheck D, Mai-Prochnow A, Burke C, Penesyan A, Dalisay DS, Stelzer-Braid S, Saunders N, Johnson J, Ferriera S, Kjelleberg S, Egan S**. 2008. Analysis of the *Pseudoalteromonas tunicata* genome reveals properties of a surface-associated life style in the Marine environment. PLoS One **3:**e3252. http://dx.doi.org/10.1371/journal.pone.0003252.
13. **Wolf T, Shelest V, Nath N, Shelest E**. 2015. CASSIS and SMIPS—promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. Bioinformatics, in press.
14. **Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH**. 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res **43:**W237–W243. http://dx.doi.org/10.1093/nar/gkv437.

2.6 Manuskript 5: »Draft Genome Sequences of Six *Pseudoalteromonas* Strains, P1-7a, P1-9, P1-13-1a, P1-16- 1b, P1-25, and P1-26, Which Induce Larval Settlement and Metamorphosis in *Hydractinia echinata*«

Status veröffentlicht, Dezember 2015

Literaturangabe Klassen, Jonathan L. ; Wolf, Thomas ; Rischer, Maja ; Guo, Huijuan ; Shelest, Ekaterina ; Clardy, Jon ; Beemelmanns, Christine: Draft Genome Sequences of Six *Pseudoalteromonas* Strains, P1-7a, P1-9, P1-13-1a, P1-16- 1b, P1-25, and P1-26, Which Induce Larval Settlement and Metamorphosis in *Hydractinia echinata*. In: *Genome Announcements* 3 (2015), Dezember, Nr. 6, e01477–15. `http://dx.doi.org/10.1128/genomeA.01477-15`. – DOI 10.1128/genomeA.01477–15. – ISSN 2169–8287

Übersicht Das Bakterium *Pseudoalteromonas* lebt oft im Verbund mit marinen wirbellosen Tieren. Sechs *Pseudoalteromonas*-Stämme wurden von der marinen Hydrozoe *Hydractinia echinata* isoliert. Der Einfluss von *Pseudoalteromonas* auf die Anhaftung der Larven an Oberflächen und die Entwicklung von *Hydractinia echinata* wurde untersucht. Außerdem wurden die Genome der *Pseudoalteromonas*-Stämme sequenziert und annotiert. Es wurden Gene identifiziert, die im Zusammenhang mit Sekretion, Biofilmbildung und der Produktion von Sekundärmetaboliten (genomweite Suche nach PKSs, NRPSs und DMATSs) stehen. Weiterhin wird die Genomsequenz dabei helfen, mögliche Mechanismen für die Anhaftung und Metamorphose des Polypen *Hydractinia echinata*, und anderen marinen Wirbellosen, zu verstehen.

Beiträge JC und CB konzipierten, planten und organisierten die Experimente. MR, HJ und CB führten die Experimente durch. TW und JK waren für die Datenanalyse zuständig. TW analysierte die Sequenzdaten in Hinblick auf die Biosynthese von Sekundärmetaboliten und die Genomkomposition. JC, ES und CB stellten Material und Analysewerkzeuge zur Verfügung. MR und CB schrieben die Publikation.

genomeA
Journals.ASM.org

Genome Sequences of Three *Pseudoalteromonas* Strains (P1-8, P1-11, and P1-30), Isolated from the Marine Hydroid *Hydractinia echinata*

Jonathan L. Klassen,[a] Maja Rischer,[b] Thomas Wolf,[b] Huijuan Guo,[b] Ekaterina Shelest,[b] Jon Clardy,[c] Christine Beemelmanns[b]

Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut, USA[a]; Leibniz Institute for Natural Product Research and Infection Biology e.V., Jena, Germany[b]; Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA[c]

The genomes of three *Pseudoalteromonas* strains (P1-8, P1-11, and P1-30) were sequenced and assembled. These genomes will inform future study of the genes responsible for the production of biologically active compounds responsible for these strains' antimicrobial, biofouling, and algicidal activities.

Received 5 October 2015 **Accepted** 23 October 2015 **Published** 10 December 2015

Citation Klassen JL, Rischer M, Wolf T, Guo H, Shelest E, Clardy J, Beemelmanns C. 2015. Genome sequences of three *Pseudoalteromonas* strains (P1-8, P1-11, and P1-30), isolated from the marine hydroid *Hydractinia echinata*. Genome Announc 3(6):e01380-15. doi:10.1128/genomeA.01380-15.

Address correspondence to Christine Beemelmanns, christine.beemelmanns@hki-jena.de.

Marine pseudoalteromonads are commonly associated with diverse marine eukaryotic hosts (1, 2) and exhibit a remarkable ability to produce small molecules with a broad range of bioactivities, including antibacterial (3), (anti)biofouling (4, 5), and algicidal (6) activities. We isolated three *Pseudoalteromonas* strains from the tissue of *Hydractinia echinata*, a colonial marine hydroid growing on gastropod shells inhabited by hermit crabs (*Pagurus pollicaris*). Sequencing these strains' genomes will assist the manipulation of *Pseudoalteromonas* genomes, facilitate the discovery and production of new and biologically active molecules (7), and might provide insights into the molecular cues and mechanisms involved in the recruitment and settlement of *H. echinata* larvae (8).

Freshly collected *H. echinata* were purchased from the Marine Biological Laboratory (Woods Hole, MA, USA), and the tissue surface of feeding polyps were investigated for the presence of bacteria from the *Pseudoalteromonas* genus. Clean isolates were cultured in marine broth (Difco 2216) for 3 days at 30°C (150 rpm), and metabolites were extracted using standard solid-phase extraction methods. The resulting organic extracts were tested for antimicrobial activity against a broad range of human pathogenic bacteria and fungi, and showed weak to moderate antimicrobial activity against Gram-positive bacteria (e.g., *Staphylococcus aureus*). Genomic DNA was extracted using the GenElute blood genomic DNA kit (Sigma-Aldrich) according to the manufacturer's protocol. Sequencing was performed at the Harvard Medical School Biopolymers Facility using Illumina TruSeq 50-bp paired-end libraries and a HiSeq2000 instrument (Illumina CASAVA version 1.8.2). A fraction of these reads representing ~50× coverage were assembled using the A5 pipeline version 201401013 (9) and screened for potential contaminations using blobology (10). Genomes were annotated using Prokka version 1.10 (11), and statistics were calculated using scripts from the Assemblathon 2 project (12).

The draft genome of strain P1-8 was sequenced to 50× coverage and comprises 37 contigs in 29 scaffolds, totaling 4,488,653 bases in length and having a G+C content of 41.2%. Its annotation includes 3,992 coding sequences (CDSs), 36 tRNAs, and 3 rRNAs.

The draft genome of strain P1-11 was sequenced to 51× coverage and comprises 44 contigs in 31 scaffolds, totaling 4,377,754 bases in length and having a G+C content of 41.0%. Its annotation includes 3,885 CDSs, 39 tRNAs, and 3 rRNAs.

The draft genome of strain P1-30 was sequenced to 51× coverage and comprises 51 contigs in 35 scaffolds, totaling 4,337,278 bases in length and having a G+C content of 40.9%. Its annotation includes 3,824 CDSs, 36 tRNAs, and 3 rRNAs.

Genes associated with biofilm formation and surface attachments, including genes encoding for *curli*, type II secretion system, type IV pili, and capsular polysaccharide (O-antigen) were identified, reflecting the adaptation to successful persistence and competition on marine surfaces (13). Genes encoding for secondary metabolite production (e.g., alterochromides), bacteriocins, and siderophore function (e.g., desferrioxamines) were detected using antiSMASH (14) and SMIPS (15). These genomes will promote the genetic analysis of the *Pseudoalteromonas* genus and will provide insights into secondary metabolite production and the molecular cues and mechanisms involved in the recruitment and settlement of *H. echinata* larvae (8).

Nucleotide sequence accession numbers. The whole-genome shotgun projects for strains P1-8, P1-11, and P1-30 have been deposited in DDBJ/EMBL/GenBank under the accession numbers LJSO00000000, LJSP00000000, and LKBC00000000, respectively. The versions described in this paper are the first versions, LJSO01000000, LISP01000000, and LKBC01000000.

ACKNOWLEDGMENTS

We are grateful for financial support from the NIH to J.C. (GM086258) and the German National Academy of Sciences Leopoldina for a postdoctoral fellowship to C.B. (LPDS 2011-2). J.L.K. was supported by funds from the University of Connecticut. M.R. was supported by the graduate school Jena School for Microbial Communication (JSMC), financed by the Deutsche Forschungsgemeinschaft, and T.W. was supported by the

International Leibniz Research School for Microbial and Molecular Interactions (ILRS), as part of the JSMC.

REFERENCES

1. **Gauthier G, Gauthier M, Christen R**. 1995. Phylogenetic analysis of the genera *Alteromonas*, *Shewanella*, and *Moritella* using genes coding for small-subunit rRNA sequences and division of the genus *Alteromonas* into two genera, *Alteromonas* (emended) and *Pseudoalteromonas* gen. nov., and proposal of twelve new species combinations. Int J Syst Bacteriol **45:**755–761. http://dx.doi.org/10.1099/00207713-45-4-755.
2. **Holmström C, Kjelleberg S**. 1999. Marine *Pseudoalteromonas* species are associated with higher organisms and produce biologically active extracellular agents. FEMS Microbiol Ecol **30:**285–293. http://dx.doi.org/10.1016/S0168-6496(99)00063-X.
3. **Bowman JP**. 2007. Bioactive compound synthetic capacity and ecological significance of marine bacterial genus *Pseudoalteromonas*. Mar Drugs **5:**220–241. http://dx.doi.org/10.3390/md504220.
4. **Qian P-Y, Lau SCK, Dahms H-U, Dobretsov S, Harder T**. 2007. Marine biofilms as mediators of colonization by marine macroorganisms: implications for antifouling and aquaculture. Mar Biotechnol **9:**399–410. http://dx.doi.org/10.1007/s10126-007-9001-9.
5. **Holmström C, Egan S, Franks A, McCloy S, Kjelleberg S**. 2002. Antifouling activities expressed by marine surface associated *Pseudoalteromonas* species. FEMS Microbiol Ecol **41:**47–58. http://dx.doi.org/10.1016/S0168-6496(02)00239-8.
6. **Lovejoy C, Bowman JP, Hallegraeff GM**. 1998. Algicidal effects of a novel marine *Pseudoalteromonas* isolate (class *Proteobacteria*, gamma subdivision) on harmful algal bloom species of the genera *Chattonella*, *Gymnodinium*, and *Heterosigma*. Appl Environ Microbiol **64:**2806–2813.
7. **Machado H, Sonnenschein EC, Melchiorsen J, Gram L**. 2015. Genome mining reveals unlocked bioactive potential of marine gram-negative bacteria. BMC Genomics **16:**158–170. http://dx.doi.org/10.1186/s12864-015-1365-z.
8. **Frank U, Leitz T, Müller WA**. 2001. The hydroid *Hydractinia*: a versatile, informative cnidarian representative. Bioessays **23:**963–971. http://dx.doi.org/10.1002/bies.1137.
9. **Coil D, Jospin G, Darling AE**. 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinformatics **31:**587–589. http://dx.doi.org/10.1093/bioinformatics/btu661.
10. **Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M**. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet **4:**237. http://dx.doi.org/10.3389/fgene.2013.00237.
11. **Seemann T**. 2014. Prokka: rapid prokaryotic genome annotation. BioInformatics **30:**2068–2069. http://dx.doi.org/10.1093/bioinformatics/btu153.
12. **Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou W, Corbeil J, Del Fabbro C, Docking T, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapthy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang H, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam T-W, Lavenier D, Laviolette F, Li Y, Li Z, Lio B, Liu Y, Luo R, MacCallum I, MacManes MD, Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu S-M, Yuan J, Zhang G, Zhang H, Zhou S, Korf IF**. 2013. Assemblathon 2: Evaluating *de novo* methods of genome assembly in three vertebrate species. GigaScience **2:**10. http://dx.doi.org/10.1186/2047-217X-2-10.
13. **Thomas T, Evans FF, Schleheck D, Mai-Prochnow A, Burke C, Penesyan A, Dalisay DS, Stelzer-Braid S, Saunders N, Johnson J, Ferriera S, Kjelleberg S, Egan S**. 2008. Analysis of the *Pseudoalteromonas tunicata* genome reveals properties of a surface-associated life style in the marine environment. PLoS One **3:**e3252. http://dx.doi.org/10.1371/journal.pone.0003252.
14. **Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH**. 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res **43:**W237–W243. http://dx.doi.org/10.1093/nar/gkv437.
15. **Wolf T, Shelest V, Nath N, Shelest E**. 2015. CASSIS and SMIPS—promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. Manuscript under revision. Program available at https://sbi.hki-jena.de/smips.

2.7 Manuskript 6: »Genetic and metabolic aspects of primary and secondary metabolism of the Zygomycetes«

Status veröffentlicht, Februar 2016

Literaturangabe Voigt, Kerstin ; Wolf, Thomas ; Ochsenreiter, Katrin ; Nagy, Gábor ; Kaerger, Kerstin ; Shelest, Ekaterina ; Papp, Tamás: 15 Genetic and Metabolic Aspects of Primary and Secondary Metabolism of the Zygomycetes. In: *Biochemistry and Molecular Biology*. Springer Cham, 2016 (The Mycota). `https://link.springer.com/chapter/10.1007/978-3-319-27790-5_15`. – ISBN 978–3–319–27788–2 978–3–319–27790–5, 361–385. – DOI 10.1007/978–3–319–27790–5_15

Übersicht Zygomyzeten sind die ältesten terrestrischen Pilze. Die Entwicklung ihres Primär- uns Sekundärmetabolismus ist geprägt durch die frühe Koevolution mit anderen landlebenden Mikroorganismen. Neueste Genomanalysen deuten auf ein bisher unbekanntes Potential zur Herstellung von Sekundärmetaboliten hin. Dieser Übersichtsartikel weist auf die besondere Bedeutung der Zygomyzeten hin. Beschrieben werden vor allem verschiedene Aspekte ihres Metabolismus und dessen Regulation, sowie die Möglichkeit zur Gewinnung von neuartigen Wirkstoffen. Um das Potential zur Herstellung von Sekundärmetaboliten abschätzen zu können, wurden alle verfügbaren Zygomyzetengenome auf PKS-, NRPS- und DMATS-Gene hin untersucht.

Beiträge KV und ES konzipierten und planten den Übersichtsartikel. TW analysierte die Genomsequenzdaten in Hinblick auf NRPSs, PKSs/FASs, DMATSs und deren Verteilung in den Zygomyzetengenomen. KK analysierte die Evolution von Karotinoiden und stellte eine entsprechende Abbildung zur Verfügung. KV, ES, KO, GN und TP schrieben die Teile der Publikation. KV übernahm das abschließende Layout und die Korrespondenz.

[2]Mit Genehmigung durch *Springer Nature*

15 Genetic and Metabolic Aspects of Primary and Secondary Metabolism of the Zygomycetes

Kerstin Voigt[1,2], Thomas Wolf[3], Katrin Ochsenreiter[4], Gábor Nagy[5], Kerstin Kaerger[6], Ekaterina Shelest[3], Tamás Papp[5]

CONTENTS

I. Introduction 361
A. Zygomycetes: Evolution, Systematics, and Ecology 362
B. The Cooperative Nature of Zygomycetes: Bacterial–Fungal Alliances 363
II. Key Aspects in the Metabolism of Zygomycetes: Biotechnological Implications 364
A. Carotene Biosynthesis and Degradation: Primary Meets Secondary Metabolism 365
1. Regulation, Genetic Manipulation: What Have We Learned from the Major Model Organisms *Mucor circinelloides*, *Phycomyces blakesleeanus*, and *Blakeslea trispora*? 367
2. Carotene Degradation Is Linked to Sexual Interactions 370
B. Fatty Acids 372
C. Organic Acids 372
D. Storage Lipids and Single Cell Oils 373
E. Enzymes 374
III. The Dogma of the Unability of Zygomycetes to Produce Natural Products 375
IV. Conclusions 377
References 377

[1]Jena Microbial Resource Collection, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute, Adolf-Reichwein-Strasse 23, 07745 Jena, Germany; e-mail: kerstin.voigt@leibniz-hki.de
[2]Department of Microbiology and Molecular Biology, Institute of Microbiology, University of Jena, Neugasse 25, 07743 Jena, Germany; e-mail: kerstin.voigt@leibniz-hki.de
[3]Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute, Adolf-Reichwein-Strasse 23, 07745 Jena, Germany
[4]Karlsruhe Institute of Technology (KIT), Institute of Process Engineering in Life Sciences, Section II: Technical Biology, Engler-Bunte-Ring 1, 76131 Karlsruhe, Germany
[5]Faculty of Science and Informatics, Department of Microbiology, University of Szeged, Közép fasor 52, 6726 Szeged, Hungary
[6]National Reference Center for Invasive Mycoses, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute, Adolf-Reichwein-Strasse 23, 07745 Jena, Germany

I. Introduction

Zygomycetes constitute a remarkable group of microscopic fungi (formerly classified into the phylum *Zygomycota*) basal to *Ascomycota* and *Basidiomycota* (for review see Voigt 2012; Voigt and Kirk 2014). These fungi are mainly soil inhabitants living as saprobes and decomposers of organic matter and herbivorous feces (coprophiles). Some taxa are parasitic or predacious, in which case developing mycelium is immersed in the host tissue:

Traditionally, the Zygomycota, represent the most basal terrestrial phylum of the kingdom of Fungi. The Zygomycota are not accepted as a valid phylum (as "Phylum des Zygomycètes"; Whittaker 1969; Cavalier-Smith 1981 because of a lacking compliance to the International Code of Botanical Nomenclature/International Code of Nomenclature for algae, fungi and plants (Hawksworth 2011) and lacking resolution of the basal fungal clades (James et al. 2006). Molecular phylogenetic analyses based on informal phylogenetic trees where molecular phylogenies are substituted with traditional taxonomic information revealed dispersal into five subphyla containing one to four orders (Hibbett et al. 2007; Hoffmann et al. 2011, for review see: Benny et al. 2014). The phylogenetic relationships between these subphyla and their orders is still not well resolved. However, based on the potential of all five subphyla to produce zygospores during conjugation of two yoke-shaped gametangia it is referred to a phylogenetically coherent group named zygosporic fungi as a whole group, which share morphological features but consists of phylogenetically unrelated subphyla. Therefore, the

Biochemistry and Molecular Biology, 3rd Edition
The Mycota III
D. Hoffmeister (Ed.)

phylum referred to as "Zygomycota" is employed to make clear the term is being used in a colloquial sense, for instance the inclusion of all basal lineages of terrestrial fungi with the potential to form zygospores or sharing any other of the plesiomorphic morphological characters of the former phylum.

Cavalier-Smith (1981, 1998) introduced the name as "cl. nov." and comments that it does not appear to have been validly published elsewhere. Likewise, this class *Zygomycetes* does not appear to be monophyletic (James et al. 2006, for review and comprehensive phylogeny see Voigt and de Hoog 2013). Here the term "zygomycetes" is printed in lower-case letters and used in a colloquial sense for ecological groupings sharing soil as their main habitat. Zygomycetes are common and cosmopolitan components of the mycoflora of dung, soil, and other substrates that support their growth and sporulation.

A. Zygomycetes: Evolution, Systematics, and Ecology

Zygomycetes are ubiquitously distributed. The occurrence of zygomycetes dates back to the Precambrian era, 800–1400 million years ago (Heckman et al. 2001; Mendoza et al. 2014). During the course of their long evolutions, which has its roots in the Precambrian, they have learned to interact with many other microorganisms in a wide variety of interplays, e.g., symbiosis with endo- and ectosymbionts; commensalism and parasitism during zygomycetes become successful pathogens of plants, animals, and human. Interaction pattern appears to be instrumental for fitness reasons as shown in aphid–bacterium–fungus alliances lowering the rate of transmission diseases (Scarborough et al. 2005).

Zygomycetes encompass nine orders: *Asellariales*, *Dimargaritales*, *Endogonales*, *Entomophthorales*, *Harpellales*, *Kickxellales*, *Mortierellales*, *Mucorales*, and *Zoopagales* (for review see Voigt and Kirk 2014). Members of the *Asellariales* (18 species) have filamentous, branched thalli and reproduce asexually by arthrospore-like cells that disarticulate from their corresponding thallus. They inhabit the digestive tract of terrestrial, aquatic, and marine isopods as well as springtails by attachment to the cuticle or digestive tract via a holdfast. They are not immersed in the host tissue (Moss 1975; Lichtwardt and Manier 1978). The *Dimargaritales* (18 species) is comprised by **obligate haustorial mycoparasites of the** *Mucorales* (rarely, species of *Chaetomium* [*Ascomycota*: *Sordariales*]) which are saprobic, or coprophilous, and share the same habitat.

The *Endogonales* (15 species) is an order of **mainly ectomycorrhizal fungi**, in addition to some saprobes. Endogonalean fungi are widely associated with the earliest branching land plants. During their evolution, they give way to arbuscular mycorrhizal glomeromycotan fungi in later lineages. It has been hypothesized that *Endogone*-like fungi rather than (as previously proposed, Simon et al. 1993; Parniske 2008) the *Glomeromycota* enabled the establishment and growth of early land colonists and thus facilitated terrestrialization (Bidartondo et al. 2011).

The *Entomophthorales* (250 species) consists of **mainly entomogenous/entomopathogenic fungi** producing one of the most spectacular insect-killing mechanisms. They are occasionally saprobic and found in soil, but mainly parasites of insects ("insect destroyers"), and other arthropods, rarely of nematodes and tardigrades as hosts. Most species are obligate parasites, and, therefore, these are so highly adapted to their hosts that their lifestyle obligately relies on the close relation to the host insect throughout the entire fungal ontogeny making a fungal cultivation in axenic cultures impossible. Few recipes of pure cultures have been reported which are highly complex media often containing natural products or biopolymers (Grundschober et al. 1998; Delalibera et al. 2003). Even under these conditions, it is unlikely that growth will be typical, and certainly sporulation will rarely be present. The exceptions to this general property are species of the genus *Conidiobolus* which are saprobes from the soil and are of widespread distribution. They are frequently isolated from the soil and are easy to grow in culture. *Conidiobolus coronatus* is found to be associated with medical and veterinary cases of mainly local, chronically lapsed, entomophthoromycotic infections (for review see Rothhardt et al.

2011; Mendoza et al. 2014). Zygospores, where known, are formed on differentiated hyphae.

The ecology of the *Harpellales* (252 species) is similar to that of the *Asellariales* by endocommensal association with the aquatic larvae of arthropods (incl. crustaceans and diplopods, rarely isopods), found attached to the gut lining of the aquatic larvae.

The *Kickxellales* (37 species) comprise mainly **saprobes from soil or coprophilous in dung**, rarely as mycoparasites. The *Kickxellales* are of widespread occurrence apparently favoring somewhat dry climates rather than the wet tropics but are relatively under-recorded, so their true distribution, like that for many of the fungi, is unclear.

The order *Mortierellales* (79 species) possesses an extremely high ecological and physiological diversity enabling them to be distributed worldwide (for comprehensive phylogeny, see Nagy et al. 2011; Wagner et al. 2013). Most species are polyunsaturated fatty acid-based lipid-accumulating organisms (e.g., *Mortierella alpina*, for overview see Münchberg et al. 2012, 2015). One thermotolerant species, *M. wolfii*, has clinical relevance and appears as a causative agent of bovine abortion (Papp et al. 2011). Zygospores are mostly thin walled, not ornamented, and nonpigmented.

The order *Mucorales* (237 species) is the most prominent and the most studied group among the zygomycetes (Voigt and Kirk 2014). Members of the *Mucorales* constitute a remarkable group which encompass a wide variety of morphological appearances, ecological niches, and lifestyles (saprobic, facultative parasitic, opportunistic pathogenic) facilitating extensive evolutionary studies (Voigt and Wöstemeyer 2001; Voigt et al. 2009, 2013; Hoffmann et al. 2009, 2013).

Mucoralean species are **predominantly saprotrophic, soil inhabitants**, rarely mycoparasites (biotrophic fusion parasites) on other mucoralean hosts. Due to airborne spores, high germination, and growth rates, mucoralean species belong to the primary colonizers of organic substrates. As typical indoor contaminants and post-harvest pathogens on fruits and food causing food spoilage, mucoralean fungi are the most successful and most abundant zygosporic fungi encountering permanent presence in the human environment. Some mucoralean species are able to develop life-threatening infections within immunocompromised patients (mucormycosis) (de Hoog et al. 2000, 2014; Mendoza et al. 2014; Ibrahim 2011; Chayakulkeeree et al. 2006; Greenberg et al. 2004; Bitar et al. 2009; Chakrabarti et al. 2008, 2009; Morace and Borghi 2012; Casadevall and Pirofski 2001). On the other hand, mucoralean species are used for fermentation of soy-based food in Asia since centuries and for the application of *Rhizopus* species in biotechnological production of enzymes for decades.

The *Zoopagales* (208 species) is, even though species rich, relatively unknown concerning its frequency and distribution. They appear to be cosmopolitan as obligate haustorial parasites of fungi and animals (nematodes, Amoeba, and other small terrestrial invertebrates).

B. The Cooperative Nature of Zygomycetes: Bacterial–Fungal Alliances

The observation that progressive coupling of fungal host and bacterial endosymbiont metabolic and reproductive interests leads to an acceleration of studies reporting the cooperative nature of bacterial-fungal alliances in zygomycetes (Partida-Martinez and Hertweck 2005). The macrocyclic polyketide metabolite rhizoxin has been frequently isolated from cultures of *Rhizopus microsporus*, which is infamous for causing rice seedling blight (Tsuruo et al. 1986; White et al. 2002: *R. chinensis* synonym of *R. microsporus*; Dolatabadi et al. 2014a). Among other antimicrotubule agents, **rhizoxin** was proven to be particularly effective in small-cell lung cancer cell lines with a potential application in the salvage treatment of refractory or relapsed patients suffering small-cell lung cancer to overcome drug-resistance (Ikubo et al. 1999). **Rhizoxin is not biosynthesized by the fungus itself** but by an endosymbiotic, that is, intracellular living, bacterium of the genus *Burkholderia* (Partida-Martinez and Hertweck 2005). The remarkably complex symbiotic–pathogenic relationship that extends the fungus–plant interaction to a third, bacterial, key player unveils new perspectives for pest con-

trol. This finding appeared to be initially unexpected and unique, but the cases of endosymbiotic bacterial alliances with zygomycetes have increased during the following time (*Mortierella elongata*: Sato et al. 2010; Bonito et al. 2013, *Rhizopus chinensis*: White et al. 2002). All bacterial endosymbionts discovered so far in the zygomycetes belong to the family *Burkholderiaceae* (class *Betaproteobacteria*, Sato et al. 2010) and are closely related to *Glomeribacter gigasporarum*, which is an obligate endosymbiotic bacterium of the arbuscular mycorrhizal fungus *Gigaspora margarita* (Bianciotto et al. 2003). *Glomeribacter gigasporarum* reveals an interphylum network of nutritional interactions (Ghignone et al. 2012). On the other hand, the ~2.6 MB endosymbiont genome of *M. elongata* is larger than that of *Glomeribacter* but reduced compared to free-living *Burkholderia* (Bonito et al. 2013; Fujimura et al. 2014). Thus, intimate coevolution seems to be more recent than that of the alliance between *Glomeribacter gigasporarum* and *Gigaspora margarita*. Although many genes have been lost (e.g., genes encoding starch- or sucrose-degradation enzymes, phosphofructokinase leading to an incomplete glycolysis pathway, enzymes involved in the synthesis of the essential amino acids arginine, isoleucine, leucine, methionine, phenylalanine, tryptophan, histidine, and valine), some gene families have expanded including those involved in protein metabolism and electron transport (e.g., genes encoding amino acid transporters such as proteins involved in phosphate, zinc, and putrescine uptake; Ghignone et al. 2012). A gene cluster coding for a dipeptide/heme/δ-aminolevulinic acid transporters (*dpp* operon) contains the *dppA* gene, the product of which is responsible for the specificity of the imported oligopeptides and is present in at least 20 copies, suggesting that peptide uptake is crucial for bacterial cell function (Ghignone et al. 2012):

Rhizopus species appear to be trans-kingdom pathogens causing soil-, air and foodborne diseases in plants and humans (de Hoog et al. 2000; Dolatabadi et al. 2014b). The frequency of opportunistic mycoses in human began to rise since the mid 1990s (Ribes et al. 2000; Kauffman 2004; Chamilos et al. 2007). With regard to human pathogens, endosymbiotic toxin-producing bacteria in clinical *Rhizopus* isolates appear to be rather an exception then a general feature (Partida-Martinez et al. 2008). No evidence was found that bacterial endosymbionts and rhizoxin contribute to the pathogenesis of mucormycosis (Ibrahim et al. 2008). Consequently, it remains unclear if the paradigm of modulation of virulence of opportunistic fungi by widespread use of antibacterials can be applied (Chamilos et al. 2007).

II. Key Aspects in the Metabolism of Zygomycetes: Biotechnological Implications

Since centuries zygomycetes are traditionally used for the fermentative production of food in China and Southeast Asia, e.g., for tempeh or tofu (Wikandari et al. 2012; Hesseltine 1983). However, recent studies of the last few years have shown that mucoralean species can also produce a large amount of interesting and biotechnological relevant metabolites, including organic acid, e.g., lactic acid and fumaric acid, biofuels, e.g., ethanol and biodiesel, polyunsaturated fatty acids, carotenoids, chitosan, and various enzymes, e.g., amylases, cellulases, steroid 11α-hydroxylases, phytases, proteases, and lipases. Additionally, biomass can be used as animal and fish feed due to its high nutritional value (Ferreira et al. 2013; Karimi and Zamani 2013; Meussen et al. 2012).

Zygomycetous fungi show several characteristics which are advantageous in biotechnological applications: (1) one of the **highest fungal growth rates**, enabling fast biomass accumulation; (2) ability of growing at **higher temperature** for many of the species; (3) **dimorphism** of various genera, transition from filamentous growth to yeastlike growth under oxygen limitation or at high glucose concentrations (e.g., Orlowsky 1991); (4) simple demands on culture conditions; and (5) ability to produce a high diversity of enzymes enabling growth on diverse substrates, like starch or starch-containing residual materials, lignocellulosic substrates or whey, within wide temperature and pH ranges (Zhang et al. 2007; Dyal et al. 2005; Millati et al. 2005; Sautour et al. 2002; Nahas 1988; Sajbidor et al. 1988). Mucor-

alean species are amylase positive and are able to use pentoses; therefore, they can be directly applied to ethanol production from starch-containing or lignocellulosic substrates (SSF—simultaneous saccharification and fermentation) (Deng et al. 2012; Zhang et al. 2007; Jin et al. 2005). Nevertheless, the number of mucoralean species applied to established biotechnological processes is scarce:

> Currently, only few species have been fully characterized regarding their potential to produce metabolites and enzymes. Research is primarily focussed on three genera: *Rhizopus* species for the production of organic acids, *Mucor* species for the production of ethanol and single cell oil (Ferreira et al. 2013) and *Cunninghamella* species for single cell oil production (Fakas et al. 2009). Interestingly, the transition from filamentous growth to yeast-like growth in dimorphic species is triggered by similar conditions favourable for organic acid and ethanol production, particularly high glucose concentrations along with elevated CO_2 contents (Lennartsson et al. 2009; Sharifia et al. 2008; Wolff and Arnau 2002; Serrano et al. 2001; Orlowsky 1991; Bartnicki-Garcia 1968). Therefore, due to the highly beneficial characteristics and promising biotechnological potential of mucoralean species, research exploring metabolite and enzyme production is urgently needed.

Biotechnologically relevant metabolites produced by zygomycetes are ethanol, carotenoids, fatty acids, organic acids, and single cell oils (SCOs) rich in polyunsaturated fatty acids (PUFAs) which are also named storage lipids. SCOs are known for their bifunction as a supplier of functional oils, and feedstock for biodiesel production (Huang et al. 2013). Especially organic acids and single cell oils containing PUFAs have a high market value, but suitable production strains and economically efficient processes are not available. Therefore, research on these substances would imply a high impact on white biotechnological issues.

A. Carotene Biosynthesis and Degradation: Primary Meets Secondary Metabolism

Despite their negative impact on humans and agriculture, zygomycetes could also be used in a positive way, like fermentations of food and sterols or the production of additives for food, feed, or pharmaceuticals with major interest in biological production of carotenoids, e.g. zeaxanthin, lycopene, or carotene (Hesseltine 1991; Nout and Kiers 2005; Liu et al. 2012; Rodríquez-Sáiz et al. 2012; Voigt and Kirk 2014). Since animals are not able to produce carotenoids by themselves, they depend on external sources and producers like plants, microorganisms, or fungi. Carotenoids are important pigments in animals and plants serving light protection and other physiological functions, e.g., as antioxidants, chromophores in photosynthesis or photoprotection, membrane stabilizers, and precursors for vitamin A. Carotenoid biosynthesis is known to be light stimulated (Rodríguez-Ortiz et al. 2012). Within the fungi, precursors of the carotenoids originate from the mevalonate pathway and are processed via geranylgeranyl diphosphate, phytoene, lycopene to, e.g., β-carotene. Known enzymes involved in **β-carotene synthesis** (Fig. 15.1) comprise CarRA (CarRP in *Mucor circinelloides*, containing the two domains CarR (lycopene cyclase) and CarA (phytoene synthase) (Fig. 15.1; Torres-Martínez et al. 1980; Arrach et al. 2001), CarB (phytoene dehydrogenase) (Ruiz-Hidalgo et al. 1997), CarI (Roncero and Cerdá-Olmedo 1982), CarF (Mehta et al. 1997), CarC (Revuelta and Eslava 1983), and CarD (Salgado et al. 1989).

β-carotene itself is processed differently in different fungal phyla, e.g., to neurosporaxanthin in *Ascomycota* or dihydroactinidiolide and β-ionone serving as flavor compounds in the *Basidiomycota* (Zorn et al. 2003). In the basal fungal lineages, however, carotene is cleaved to pheromones facilitating sexual recognition between mating partners—the sesquiterpene sirene in the *Chytridiomycota* (*Blastocladiomycetes*, *Neocallimastigomycetes*, *Monoblepharidomycetes*, and *Chytridiomycetes* for review see Voigt et al. 2013) and trisporoids in the zygomycetes (for review see Wöstemeyer et al. 2005). Zygomycetes recruit their recognition molecules from β-carotene biosynthesis (Fig. 15.1) and degradation (Fig. 15.1) pathways bridging primary and secondary metabolism, a feature comparable to abscisic acid in plants (Schwartz et al. 1997). Trisporoids are a rather

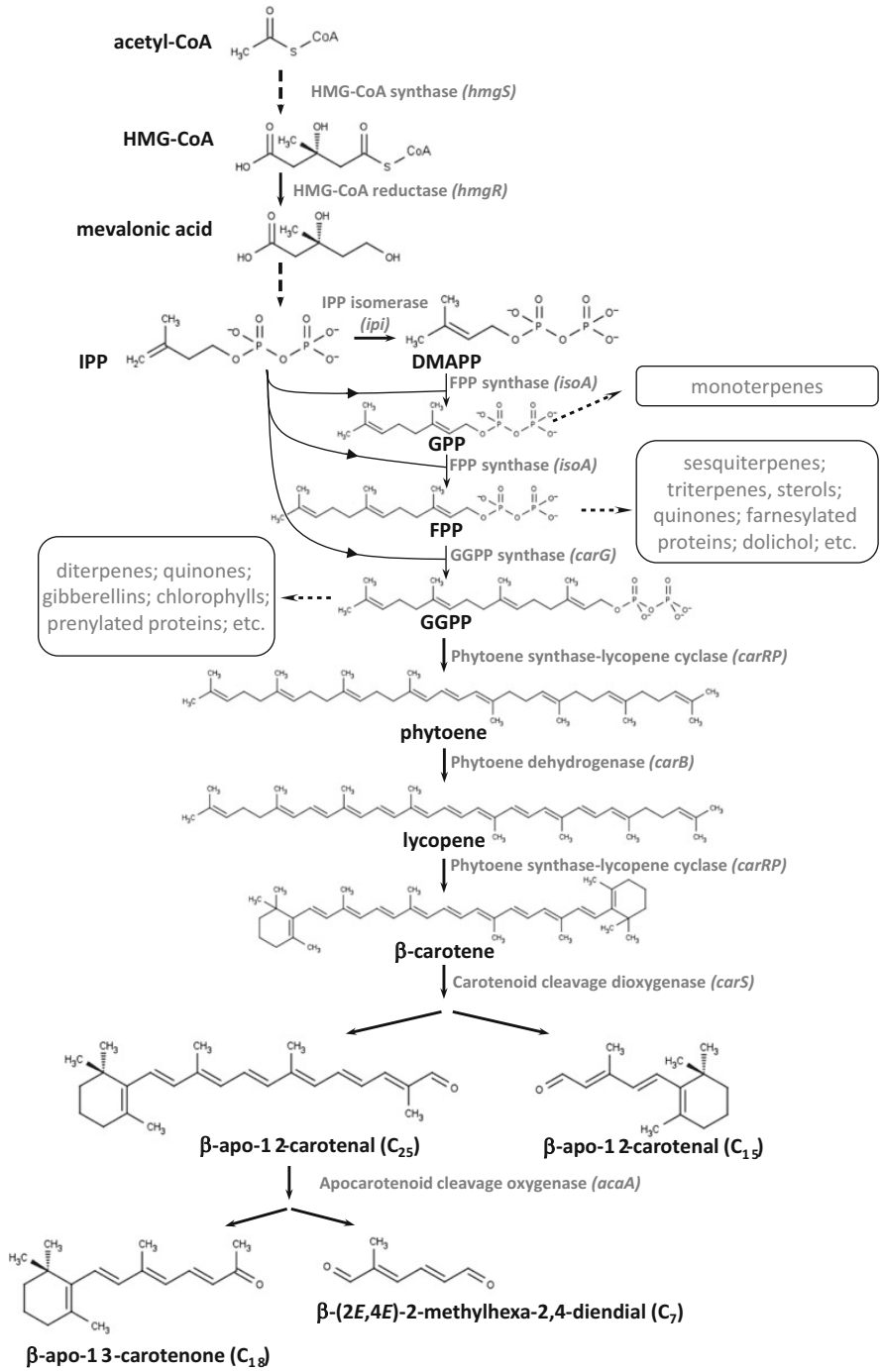

Fig. 15.1 Carotene biosynthesis and degradation pathways: Main steps of the acetate–mevalonate pathway, the specific β-carotene biosynthesis in *M. circinelloides* and presumable cleavage of β-carotene (C40)

unusual degradation product since it is not involved in cell-supporting or cell-protective functions but in pheromone action in sexual communication of those fungi if compared to higher fungi which rely on modified peptides (Gooday 1974; Jones and Bennett 2011). Since a multitude of structural diverse carotenoids is involved in a very broad and diverse spectrum of applications in all organisms, an even much more diversity of enzymes involved in carotenogenesis and processing should be reasonable. Enzymes cleaving specific double bonds are termed **carotenoid cleavage oxygenases** (CCOs) or more specified mono- (CMOs) or dioxygenases (CCDs):

Elucidation and clarification of enzymatic mechanisms started only several years ago with the first crystal structure of a CCO from *Synechocystis* sp. converting β-apo-carotenals as sole substrates (Kloer et al. 2005). Since then, only few amino acid residues are believed to be essential for enzyme activity, namely four histidine and three glutamate/aspartate residues (Poliakov et al. 2005; Takahashi et al. 2005). According to the first structurally described CCO, these amino acids correspond to sequence positions Glu150, His183, His238, His304, Glu370, Glu426 and His484 of *Synechocystis*.

All representative carotenoid cleavage enzymes (shown in Fig. 15.1) possess these conserved amino acids. In accordance with their first description by Medina et al. 2011, few sequences (clustering within the zygomycetous order *Mucorales* and termed "unknown") possess similar sequences but lack some of the conserved residues (essentially His183 and His238). The gene coding for the CCO from *Phycomyces blakesleeanus* (acaA) seems to have been duplicated recently. A functional characterization remains to be done for the duplicated *aca*A and the presumed genes of unknown function (Medina et al. 2011). The phylogeny of the CCOs (Fig. 15.2) shows that each clade has evolved its more or less specific carotenoid cleavage enzymes with similar cleaving sites, but with different natural substrates, which is presumably due to their wide variety within carotenogenesis and specific organismal requirements. The carotenoid-cleaving enzymes in the *Mucorales* are unique for this group of fungi with no similar cleaving enzymes in other fungal groups (Sahadevan et al. 2013):

Mucorales seem to possess also only one enzyme capable to cleave β-carotene, an enzyme crucial for all subsequent utilizations of β-carotene. This gene, called *carS*, is an 11′–12′ carotenoid cleavage dioxygenase (Fig. 15.1; Medina et al. 2011; Tagua et al. 2012; Rodríquez-Sáiz et al. 2012; Rodríguez-Ortiz et al. 2012), which cleaves β-carotene (C_{40}) into β-apo-12-carotenal (C_{25}) and β-apo-12-carotenal (C_{15}, Fig. 15.1, Sahadevan et al. 2013). CarS should not be misapplied as an orthologue of the *carS* in the ascomycete *Fusarium* sp., which codes for a regulatory protein, most likely corresponding to CrgA from *Mucor circinelloides* (Navarro et al. 2001). After cleavage by CarS, β-apo-12-carotenal (C_{25}) is further processed by the apocarotenoid cleavage oxygenase AcaA, presumably cleaved at its 13–14 position, resulting finally in two more fragments, namely β-apo-13-carotenone (C_{18}, also named d'orenone, Sahadevan et al. 2013), and probably (2*E*,4*E*)-2-methylhexa-2,4-diendial (C_7) (Fig. 15.1; Polaino et al. 2010; Medina et al. 2011).

1. Regulation, Genetic Manipulation: What Have We Learned from the Major Model Organisms *Mucor circinelloides*, *Phycomyces blakesleeanus*, and *Blakeslea trispora*?

Members of the order *Mucorales* are known as β-carotene-producing fungi. Among them, *Blakeslea trispora*, *Mucor circinelloides*, and *Phycomyces blakesleeanus* are involved in the study of the carotenoid biosynthesis as model organisms. *B. trispora* is already an industrial

Fig. 15.1 (continued) at position C11–C12 by the carotenoid cleavage dioxygenase CarS, resulting in two fragments of β-apo-12-carotenal, (C25) and (C15) in *P. blakesleeanus*, the final cleavage of β-apo-12-carotenal (C25) to β-apo-13-carotenone (C18) and probably (2*E*, 4*E*)-2-methylhexa-2,4-diendial (C7) by the apocarotenoid cleavage oxygenase AcaA. The most important enzymes and the encoding genes are indicated with *gray*. *HMG* hydroxymethylglutaryl, *IPP* isopentenyl pyrophosphate, *DMAPP* dimethylallyl pyrophosphate, *GPP* geranyl pyrophosphate, *FPP* farnesyl pyrophosphate, *GGPP* geranylgeranyl pyrophosphate

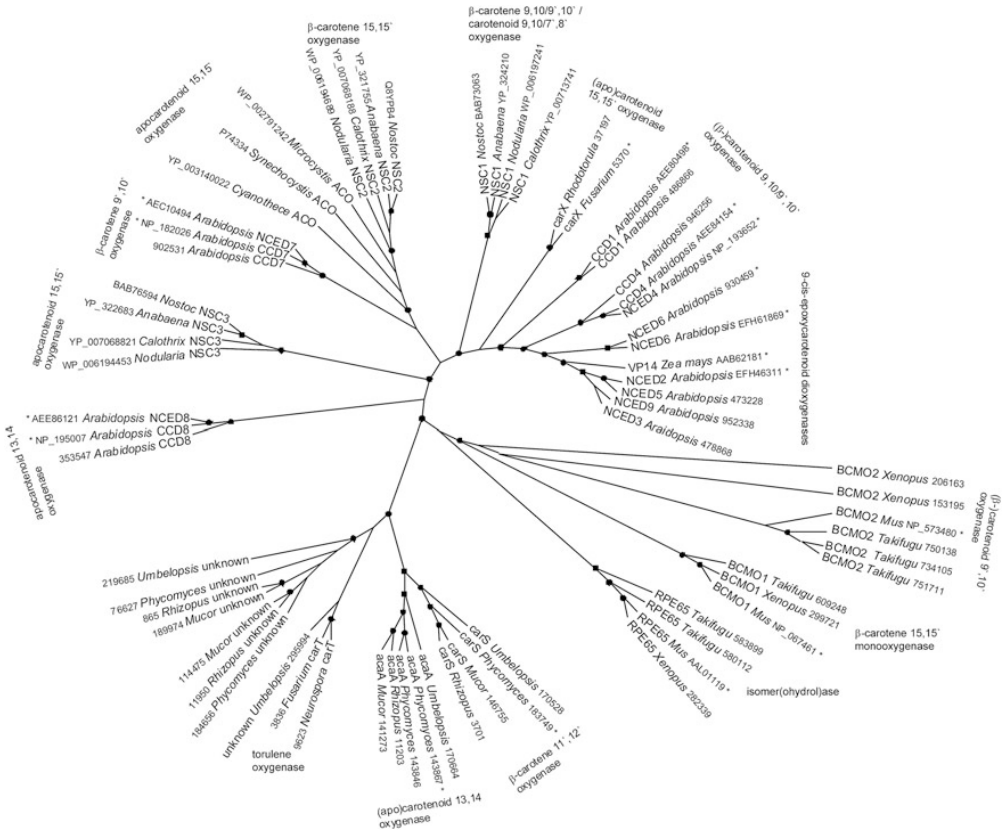

Fig. 15.2 Evolutionary relationships of representative genes involved in β-carotene degradation and their representative cleavage sites. The tree can be roughly divided into three sub-trees, each following species phylogeny. Carotenoid cleavage oxygenases from *Mucorales* comprise CarS and AcaA as well as several so far uncharacterized sequences. Bootstrap values greater or equal to 90 % are indicated by *black dots*. *Names on branches* indicate prominent cleavage sites

source of β-carotene, while the application of *M. circinelloides* and *P. blakesleeanus* is in a developmental phase (Dufossé 2006, 2008). However, improvement and study of the carotenogenesis in *B. trispora* and *P. blakesleeanus* are hampered by the lack of efficient methods for genetic manipulation; i.e., their genetic transformation has still been unsuccessful (Obraztsova et al. 2004; Sanz et al. 2011; Garre et al. 2015). *M. circinelloides* seems to be more amenable to molecular techniques as well-developed transformation systems including vectors, promoters, recipient strains, and methods (i.e., PEG-mediated protoplast transformation and electroporation) are available (van Heeswijck and Roncero 1984; Wolff and Arnau 2002; Appel et al. 2004; Papp et al. 2010; Gutiérrez et al. 2011). Moreover, this fungus has an ability to maintain and express exogenous genes from related fungi (e.g., *P. blakesleeanus*, *B. trispora*, or *Rhizomucor miehei*) and other organisms (e.g., *Xanthophyllomyces dendrorhous* or *Paracoccus* sp. N81106) (Iturriaga et al. 1992; Ruiz-Hidalgo et al. 1999; Quiles-Rosillo et al. 2003; Rodríguez-Sáiz et al. 2004; Lukács et al. 2009; Papp et al. 2006, 2013; Csernetics et al. 2015).

Carotenoids are terpenoid compounds, and their biosynthesis can be regarded as a side route of the general acetate–mevalonate (AMV) pathway, in which precursors of the different terpene derivatives are synthesized from acetyl CoA. Several genes encoding the enzymes, which catalyze the main steps of the AMV pathway and carotenoid biosynthesis, have been isolated and characterized in *M. circinelloides* (Velayos et al. 2000a, b, 2003; Csernetics et al. 2011; Nagy et al. 2014). Carotenogenic genes of *B. trispora* and *P. blakesleeanus* were also identified, and their functions were analyzed by expressing them in *M. circinelloides* (Rodríguez-Sáiz et al. 2004; Sanz et al. 2011).

One of the key enzymes of the general AMV pathway is 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) reductase, which catalyzes the formation of HMG-CoA from mevalonic acid. As HMG-CoA is a common intermediate of numerous different terpenoid compounds, such as carotenoids, ergosterol, prenyl groups of certain proteins, and ubiquinone, its formation is considered to be rate limiting for the carotenoid synthesis (Wang and Keasling 2002). *M. circinelloides* has three HMG-CoA reductase genes (*hmgR*), which respond differently to temperature and the oxygen level of the environment (Nagy et al. 2014). Among them, *hmgR2* and *hmgR3* seem to be especially involved in the carotenoid biosynthesis. Overexpression of these genes by changing their own promoter to that of the endogenous glyceraldehyde-3-phosphate dehydrogenase 1 gene (*gpd1*) and elevating their copy numbers increased the whole carotenoid content of the fungus 1.5–1.7-fold (Nagy et al. 2014).

Another important section of this pathway is the synthesis of the prenyl-chain intermediate compounds, which serve as precursors in the different specific side routes. The most important steps of this process are the isomerization of dimethylallyl pyrophosphate (DMAPP) and isopentenyl pyrophosphate (IPP) catalyzed by the IPP isomerase, the condensation of IPP and DMAPP to form geranyl pyrophosphate (GPP), and the extension of the prenyl chain by the addition of further IPP units to the carbon chain forming farnesyl and geranylgeranyl pyrophosphate (FPP and GGPP, respectively). Synthesis of GPP and FPP is catalyzed by the FPP synthase, while formation of GGPP is managed by the GGPP synthase. GGPP is the direct precursor of carotenoids as their specific biosynthesis starts with the condensation of two 20-carbon GGPP units leading to the synthesis of carotenoid phytoene (Iturriaga et al. 2000). In *M. circinelloides*, the IPP isomerase and the FPP and GGPP synthases are encoded by the *ipi*, *isoA*, and *carG* genes (Velayos et al. 2003, 2004; Csernetics et al. 2011). Overexpression of these genes significantly enhanced the carotenoid biosynthesis. In this study, the step determined by the *carG* gene proved to be the first bottleneck for carotenoid production, placing it under the control of the *Mucor gpd1* promoter resulted in a fourfold increase in the carotenoid content (Csernetics et al. 2011). Total carotenoid content of these strains was more than 2 mg/g (dry weight). Similarly, the expression of the *ipi* and the *carG* genes of *B. trispora* in an engineered, carotenoid-producing *E. coli* strain led to a twofold increase in the carotenoid production of the bacterium (Sun et al. 2012). These studies indicated that *ipi* and *carG* genes can be applied to improve the carotenoid production of mucoralean fungi. *M. circinelloides* requires light for carotenoid biosynthesis and transcription of *carG*, and the carotenoid-specific genes (i.e., *carB*-encoding phytoene dehydrogenase and *carRP*-encoding phytoene synthase–lycopene cyclase) are induced by blue light (Velayos et al. 2000a, b, 2003). White collar-1-like proteins, Mcwc-1b, and Mcwc-1c were found to be involved in the activation of the carotenogenic genes of *M. circinelloides* (Silva et al. 2006, 2008), while the protein CrgA proved to be a repressor of the carotenoid biosynthesis in *Mucor* (Navarro et al. 2001). Deletion of the *crgA* gene resulted in enhanced accumulation of carotenoids under both dark and light conditions (Navarro et al. 2001). Moreover, deletion of *crgA* could be used to increase the lycopene production of a *mutant M. circinelloides* strain achieving a lycopene content of 54 g/L (Nicolás-Molina et al. 2008). Recently,

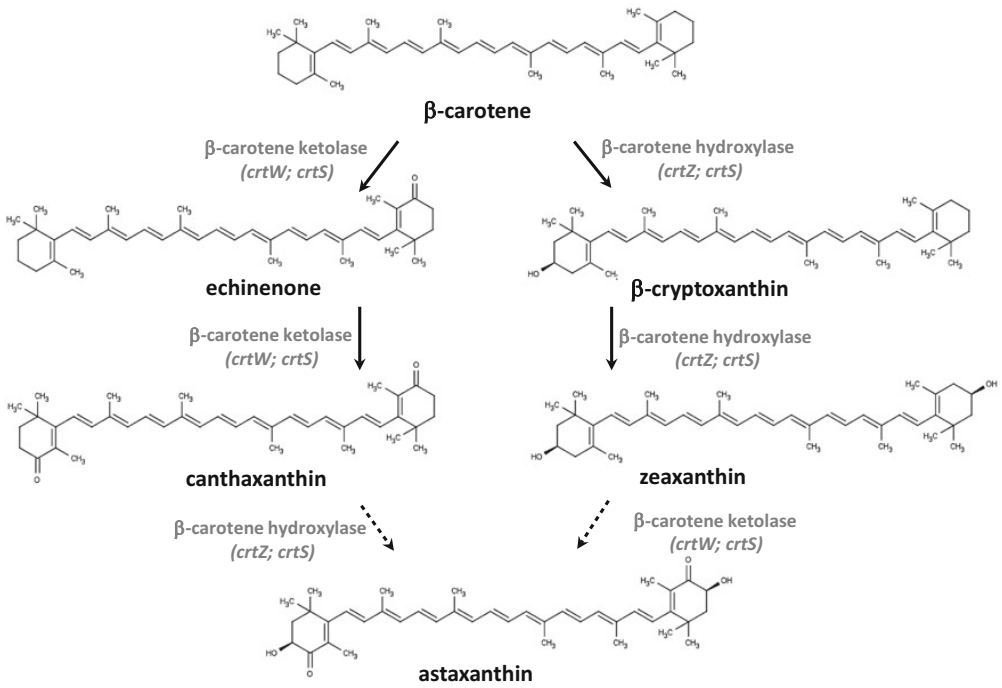

Fig. 15.3 Conversion of β-carotene to its oxygenated derivatives, the carotenoids (xanthophylls)

it has been supposed that CrgA may be an ubiquitin ligase, and one of its functions may include preventing Mcwc-1b to activate the transcription of the carotenoid biosynthesis genes (Silva et al. 2008; Navarro et al. 2013):

By expressing exogenous carotenoid biosynthesis genes, production of new carotenoid compounds, such as oxygenated derivatives of β-carotene, can be achieved. *Paracoccus* sp. N81106 is a marine, astaxanthin producing bacterium, in which the conversion of β-carotene to astaxanthin is catalyzed by the enzymes β-carotene ketolase (CrtW) and hydroxylase (CrtZ). Production of xanthophylls (i.e. β-cryptoxanthin, zeaxanthin, canthaxanthin and astaxanthin, Fig. 15.3) could be carried out by transforming *M. circinelloides* with autonomously replicating vectors, which harboured the *crtW* and the *crtZ* genes fused with the regulatory sequences of *Mucor gpd1* (Papp et al. 2006). Multiple integration of the *crtW* gene into the *Mucor* genome resulted in strains accumulating canthaxanthin as the main carotenoid instead of β-carotene (Papp et al. 2013). The astaxanthin biosynthesis gene (*crtS*) of *Xanthophyllomyces dendrorhous* also could be used to obtain xanthophyll-producing *M. circinelloides* strains (Álvarez et al. 2006; Csernetics et al. 2015; Rodríquez-Sáiz et al. 2012). In these experiments, the *crtS* gene was driven by the promoter of the *Blakeslea carRA* or the *Mucor gpd1*.

2. Carotene Degradation Is Linked to Sexual Interactions

All zygomycetes are coherently united by the potential to form the chemotactic pheromone **trisporic acid** (Fig. 15.4; for review see Wöstemeyer et al. 2002, 2005). This compound is morphogenic by its ability to induce the genesis of zygophores subsequently followed by zygospores during conjugation of two yoke-shaped gametangia (**gametangiogamy**) in compatible mating interactions. Trisporic acid is the universal gamone, which is cooperatively formed between both mating partners (Schimek et al. 2003; Schachtschabel et al. 2008). Trisporoid acid has a multitude of derivatives (trisporoids), which possess deviating biological activ-

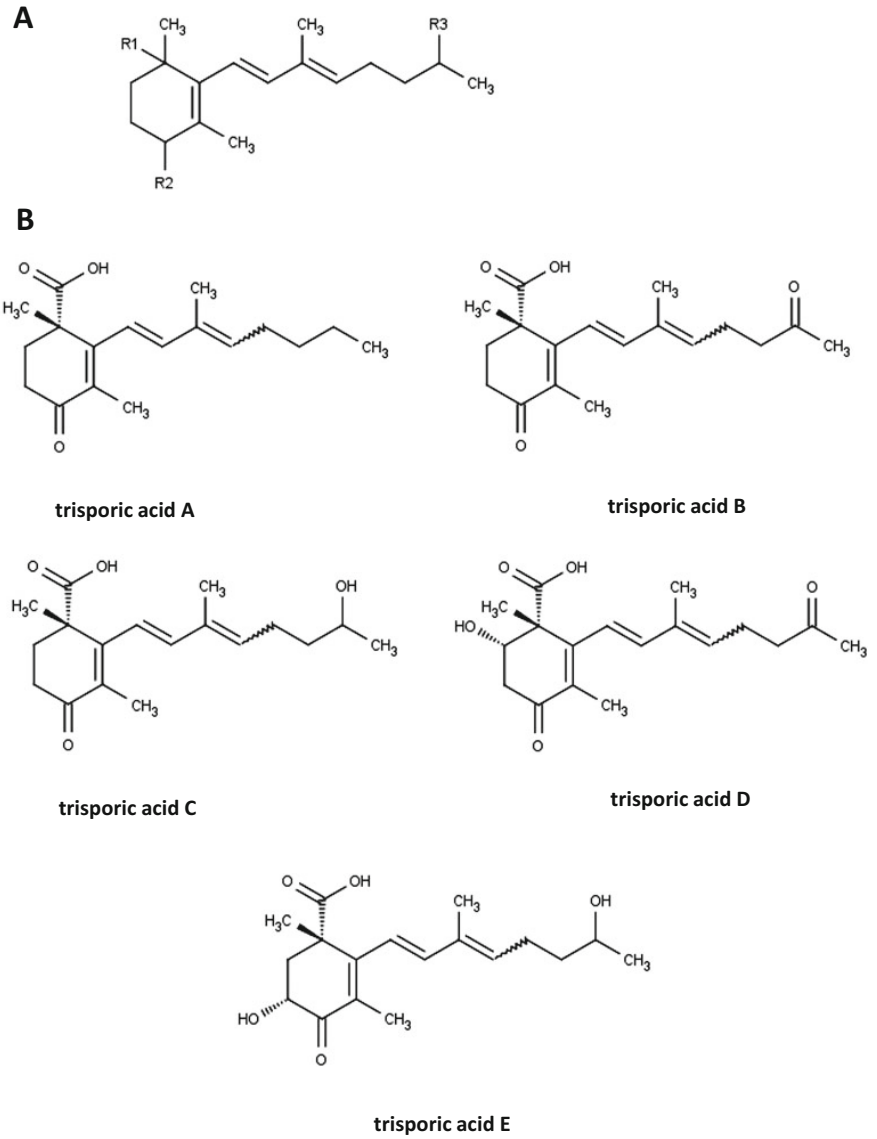

Fig. 15.4 Chemical structures of trisporic acid, the universal sexual pheromone of the zygomycetes. (**a**) Basic chemical structure. (**b**) Structures of trisporic acids A–D. Trisporic acid D was postulated but was never experimentally proven

ity (Schachtschabel et al. 2005). The biosynthesis of trisporoids starts with the cleavage of β-carotene which is mediated by a trisporic acid-regulated β-carotene oxygenases *tsp3* and *tsp4* aiming a C_{18} compound (Burmester et al. 2007). The sexual pheromones of the *Mucorales* are processed from the C_{18} compound resulting finally in trisporic acids. Yet many of the enzymatic steps remain unknown. The only enzymes known so far in the processing of the

C_{18} compound are two enzymes belonging to two different families of oxidoreductases: *tsp2*, a short-chain dehydrogenase involved in the processing of 4-dihydrotrisporin, and *tsp1*, an aldo-keto reductase involved in the processing of 4-dihydromethyltrisporate (Czempinski et al. 1996; Wetzel et al. 2009).

B. Fatty Acids

Fungi of the order *Mortierellales* and *Mucorales* have attracted considerable interest as **industrial lipid producers**. They are easily cultivated in solid or liquid culture and have been shown to grow on various different carbon sources (Dyal and Narine 2005; Gao et al. 2013; Zeng et al. 2013), on numerous different agricultural waste products (Chaudhuri et al. 1998; Jang et al. 2000; Zeng et al. 2013), and on glycerol, a by-product of biodiesel production (Hou 2008; Dedyukhina et al. 2011; Chatzifragkou et al. 2011). Hence, industrial and agricultural waste products can be converted as low-cost substrate into valuable products, providing an excellent biotechnological application for the zygomycetes. For example, lignocellulosic biomass, which is the most available and renewable source in nature, might be an ideal raw material for single cell oils (see Sect. D) production (Huang et al. 2013). Especially *Mortierella* spp. can accumulate large amounts of unusual lipids containing polyunsaturated fatty acids depending on species, strain, and growth conditions (Münchberg et al. 2012, 2015). The characterization of the genomes from oleaginous fungi like *Mortierella alpina* (Wang et al. 2011) and *M. elongata* (Bonito et al. 2013) provides insights into the genomic basis of fatty acid production. First insights into the *M. elongata* genome reveal preliminary enrichments of genes related to lipid metabolism (e.g., sphingolipids, ether lipids, and glycerophopholipids), tryptophan metabolism, siderophore group nonribosomal peptides, and glucan 1,4-alpha glucosidases compared to genome sequences of other basal fungi (Bonito et al. 2013).

C. Organic Acids

The production of relevant organic acids, namely, **L-lactic acid** and **fumaric acid**, is based on pyruvate, the end product of the glycolysis. Whereas L-lactic acid is formed directly from pyruvate by lactate dehydrogenase (Skory 2000; Pritchard 1971, 1973), fumaric acid is formed via the oxidative branch of the TCA cycle located in the cytoplasm (Fig. 15.5; Goldberg et al. 2006).

Both organic acids can be diversely applied in food industry, textile sector, cosmetic industry, and chemical and pharmaceutical industry. Lactic acid is the most abundantly produced organic acid in nature. Therefore, lactic acid production by *Rhizopus* species is a subject of intensive research and has the potential to replace the established lactic acid production processes using chemical methods or lactobacilli fermentation. When producing lactic acid by *Rhizopus* species, low-cost substrates (e.g., agricultural waste products containing any kind of plant fibers) and a wide variety of carbon sources ranging from monosaccharides to polysaccharides can be used (Guo et al. 2010; Vially et al. 2010; Yen and Lee 2010; Bulut et al. 2009; Bai et al. 2008), resulting in very high yields ranging near the theoretical maximum (Ferreira et al. 2013; Meussen et al. 2012).

Fumaric acid, a C_4-dicarboxylic acid, was identified by the US Department of Energy as one of 12 promising platform chemicals from biomass with high added value (Werpy and Petersen 2004). Presently, fumaric acid is chemically produced from crude oil and is applied in food industry as acidulant, food preservative, and flavor enhancer. Due to its bifunctionality and the double-bond fumaric acid, it is also suitable to act as polymerization starter unit for plastics or resins (Anonymus 2007; Willke and Vorlop 2004). As for lactic acid, high yields near the theoretical maximum can be achieved by microbial fermentation when using glucose as carbon source (Meussen et al. 2012; Roa Engel et al. 2008; Cao et al. 1996). Noteworthily, for each molecule of formed fumaric acid, one molecule CO_2 is fixated (Osmani and Scrutton 1985; Overman and Romano 1969).

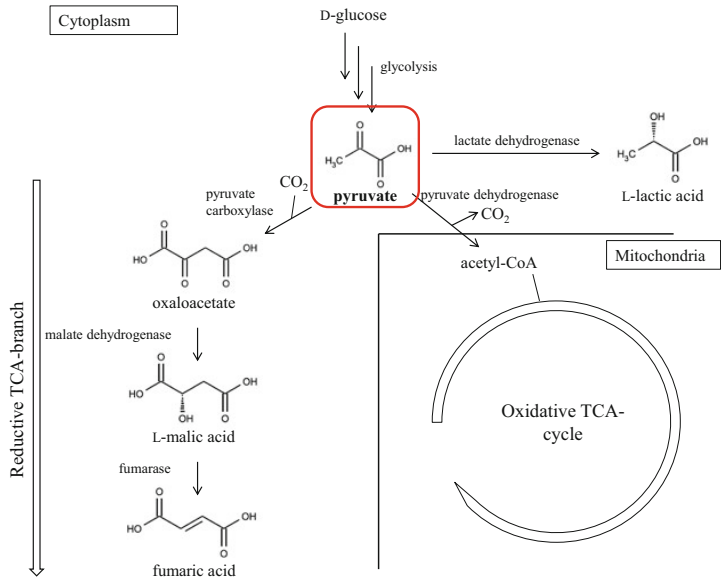

Fig. 15.5 Formation of lactic and fumaric acid is based on pyruvate, the end product of the glycolysis. Whereas L-lactic acid is formed directly from pyruvate by lactate dehydrogenase (Skory 2000; Pritchard 1971, 1973), fumaric acid is formed via the oxidative branch of the TCA cycle located in the cytoplasm

Whether fumaric or lactic acid is produced from pyruvate depends on activity and substrate affinity of the respective enzymes and seemed to be strain dependent. However, Saito et al. (2004) proved that *Rhizopus oryzae* strains with two genes for lactate dehydrogenase produce mainly lactic acid, whereas strains with only one gene produce mainly fumaric acid.

Phylogenetic studies using further independent DNA markers by Abe et al. (2007) revealed that fumaric acid and lactic acid producers can be separated into two sibling species, *R. oryzae* sensu stricto (also known as *R. arrhizus*) and *R. delemar* correlating with the lactic acid and fumaric–malic acid producers, respectively. Reclassification of strains in the fumaric–malic acid group as *R. delemar* and therefore reclassification of the genome strain *R. oryzae* 99-880 into *R. delemar* were proposed (Gryganskyi et al. 2010), which was later converted into *Rhizopus arrhizus* var. *delemar* (Dolatabadi et al. 2014a).

D. Storage Lipids and Single Cell Oils

All living organisms have to synthesize a minimum amount of lipids to build up membranes. However, only few organisms are able to accumulate more than 20 % of their dry biomass in form of storage lipids. The term “oleaginous” refers to microorganisms, including yeasts, fungi, and microalgae, which meet this criterion and store lipids in form of triacylglycerols (Ratledge and Wynn 2002). Storage lipids, which are also known as “single cell oils” (SCPs), are rich in polyunsaturated fatty acids and are of special interest due to their bifunction as a supplier of functional oils and feedstock for biodiesel production (Huang et al. 2013). Especially **γ-linoleic acid** (GLA, C18:3n−6) is biotechnologically relevant (Fig. 15.6). It is commercially applied in pharmaceutical industry and is currently obtained by extraction of selected plant oils. However, higher amounts of GLA are also produced by some mucoralean genera, like *Cunninghamella*, *Mucor* (including *Zygorhynchus*), *Rhizopus*

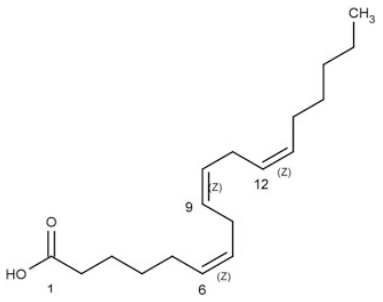

Fig. 15.6 Chemical structure of γ-linoleic acid (GLA, C18:3n−6)

(Kavadia et al. 2001; van der Westhuizen et al. 1994), *Choanephora*, *Phycomyces* (van der Westhuizen et al. 1994), and *Thamnidium* (Stredansky et al. 2000). Oleaginous microorganisms start the accumulation of single cell oil when grown in a medium with excess of carbon source but with a limitation of another nutrient. Oleagincity is characterized by the ability to produce a continuous supply of both acetyl CoA and NADPH as necessary precursors and reduction equivalent in fatty acid biosynthesis and is realized by the key enzymes ATP citrate lyase and AMP deaminase (Ratledge 2004). *Mortierellales* have great biotechnological importance as industrial producers of polyunsaturated fatty acids, such as arachidonic acid or eicosapentaenoic acid. Both the content of fatty acids and their rate of saturation are known to be dependent on the temperature during production and also vary due to utilization of different nutrients in the cultivation media (Münchberg et al. 2012, 2015).

E. Enzymes

Zygomycetes are known to produce a vast variety of enzymes, e.g., amylases, cellulases, xylanases, steroid 11α-hydroxylases, phytases, proteases, and lipases which have a multitude of applications in industrial and pharmaceutical applications (for review see Krisch et al. 2010; Voigt and Kirk 2014).

Amylases are one of the main enzymes used in industry. They hydrolyze starch molecules into polymers composed of glucose units or oligosaccharides. Amylases have potential application in industrial processes such as food, textile, paper, and detergent industries as well as fermentation and pharmaceutical industries. As starch is an important constituent of the human diet and is a major storage product of many economically important crops such as wheat, rice, maize, tapioca, and potato, starch-converting enzymes are used in the production of maltodextrin, modified starches, or glucose and fructose syrups. For the production α-amylases using submerged and solid-state fermentation systems, distribution, structural–functional aspects, physical and chemical parameters, and the use of these enzymes in industrial applications, see the review by Monteiro de Souza and de Oliveira Magalhães (2010).

Cellulases catalyze cellulolysis, the decomposition of cellulose, which is the most abundant organic source of feed/food, fuel, and chemicals (Spano et al. 1976). Cellulase breaks down the cellulose molecule into mono- and oligosaccharides by hydrolysis of the 1,4-beta-D-glycosidic linkages in cellulose in its derivative hemicellulose, lichenin, and cereal beta-D-glucans. Cellulases represent a naturally occurring mixture of various enzymes acting serially or synergistically to decompose cellulosic material. Zygomycetes (e.g., *Mucor circinelloides*) were frequently found as straw-colonizing fungi producing total cellulases, endo-beta-1,4 glucanase, and endo-beta-1,4 xylanase in solid-state fermentation (Lee et al. 2011).

Lipases are water-soluble enzymes that act on insoluble substrates and catalyze the hydrolysis of long-chain triglycerides. They play a vital role in the food, detergent, chemical, and pharmaceutical industries and have gained significant attention in the industries due to their substrate specificity and stability under varied chemical and physical conditions (for review see Gopinath et al. 2013).

Phytases are myo-inositol hexakisphosphate phosphohydrolases and represent any type of phosphatase enzyme that catalyzes the hydrolysis of phytic acid (myo-inositol hexakisphosphate)—an indigestible, organic form of phosphorus that is found in grains and oil seeds—and releases a usable form of inorganic

phosphorus (Mullaney et al. 2000). Phytases have been most commonly detected and characterized from fungi (Mullaney and Ullah 2003), specifically in the zygomycete *Rhizopus oligosporus* (DSMZ 1964), which is commonly used for tempeh production (Azeke et al. 2011). The phytases from *R. oligosporus* exhibit a broad affinity for various phosphorylated compounds. Practical interest in phytases has been stimulated by the fact that phytase supplements increase the availability of phosphorus in pig and poultry feed and thereby reduce environmental pollution due to excess phosphate excretion in areas where there is intensive livestock production.

Proteases produced by zygomycetes are rennin-like proteases secreted by several mucoralean species that are used in cheese production. In particular, mucoralean fungi (*Rhizopus oryzae*, *Circinella muscae*, *Mucor subtilissimus*, *Mucor hiemalis* f. *hiemalis*, *Syncephalastrum racemosum*, *Rhizopus microsporus* var. *chinensis*, and *Absidia cylindrospora*) were frequently isolated from maize flour, corn meal, and cooked cornflakes using surface and depth plate methods with subsequent measurement of significant proteolytic activities (de Azevedo Santiago and de Souza Motta 2008).

Steroid 11α-hydroxylases are encoded by genes of the cytochrome P450 superfamily of enzymes containing a heme cofactor (hemoproteins, Sigel et al. 2007). The cytochrome P450 proteins are monooxygenases that catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids, and other lipids. They are, in general, the terminal oxidase enzymes in electron transfer chains, broadly categorized as P450-containing systems. The term *P450* is derived from the spectrophotometric peak at the wavelength of the absorption maximum of the enzyme (450 nm) when it is in the reduced state and complexed with CO (Sigel et al. 2007). To overcome the chemically laborious stereo- and regioselective hydroxylation steps in the pharmaceutical production of corticosteroids and progestogens, zygomycetes, e.g. *Rhizopus* spp., are employed to perform the 11α-hydroxylation of the steroid skeleton, thereby significantly simplifying steroid drug production (Petrič et al. 2010).

Xylanases degrade the linear polysaccharide beta-1,4-xylan into xylose, thus breaking down hemicellulose, one of the major components of plant cell walls (Beg et al. 2001). Zygomycetes (e.g. *Mucor circinelloides*) are straw-colonizing fungi producing xylanolytic enzymes such as endo-beta-1,4 xylanase in solid-state fermentation (Lee et al. 2011).

III. The Dogma of the Unability of Zygomycetes to Produce Natural Products

It has been a widespread dogma that zygomycetes are not capable to produce own secondary metabolites, despite of those produced by endosymbiotic bacteria (see Sect. I.B.) (for examples, see Jennessen et al. 2005). However, it has been shown that zygomycetes react on other fungal secondary metabolites by morphogenic changes as shown by the sesterterpene-type phytotoxin ophiobolin produced by fungi belonging mainly to the ascomycetous genus *Bipolaris* (Krizsán et al. 2010). Ophiobolin A caused morphological changes in *Mucor circinelloides*; the fungus formed degenerated, thick or swollen cells with septa and cytoplasm effusions from the damaged cells. Here we explore the potential of zygomycetes to produce secondary metabolites together with other microorganisms in a cooperative manner.

To estimate the genomic potential of the zygomycetes to produce secondary metabolites, all final, publicly available drafts of zygomycete genomes were scanned for the presence of genes encoding **polyketide synthases** (PKSs), **nonribosomal peptide synthetases** (NRPSs), and L-**tryptophan dimethylallyl transferases** (DMATs). For more information on regulation of secondary metabolism, see also Chap. 2. A total of eight genomes were screened and analyzed: one entomophthoralean (*Conidiobolus coronatus*), one kickxellalean (*Coemansia reversa*), two mortierellalean, and four mucoralean genomes (Table 15.1). All species possess the genomic prerequisite for the production of natural products. On average, each species encodes two DMATs and 1–2 NRPSs, with a

Table 15.1 The presence of gene families encoding polyketide synthases (PKSs), nonribosomal peptide synthetases (NRPSs), dimethylallyl pyrophosphate: L-tryptophan dimethylallyl transferase (DMAT synthase) which were predicted in the genomes of 15 species comprising five subphyla of the *Zygomycota*, as of 6th of July, 2015; genome resources (if not stated elsewhere): Joint Genome Institute Broad Institute of Harvard and MIT, Origins of Multicellularity Sequencing Project *Mortierella verticillata*

		Genome resource	(PKS)/FAS[a]	NRPS	DMAT
Mucorales	*Lichtheimia hyalospora*	JGI, unpublished	1	1	2
	Mucor circinelloides	Lee et al. (2014)	2	3	2
	Rhizopus microsporus var. chinensis	Wang et al. (2013)	1(+1)	2	6
	Rhizopus arrhizus var. *delemar* (syn. *R. oryzae*)[b]	Ma et al. (2009)	1	1	3
Kickxellales	*Coemansia reversa*	Chang et al. (2015)	4	1	2
Entomophthorales	*Conidiobolus coronatus*	Chang et al. (2015)	″(1)	3	2
Mortierellales	*Mortierella verticillata*	Bonito et al. (2013)	0	1	0
	Mortierella alpina	Wang et al. (2011)[c]	1	21	0

For an overview of genome projects on basal fungi incl. *Zygomycota*, see Shelest and Voigt (2014)
[a]The genes predicted as PKS-like have the typical structure of the FAS alpha subunit
[b]Nomenclature: Gryganskyi et al. (2010) and Dolatabadi et al. (2014a)
[c]Proteome of *M. alpina* available from phylomeDB: ftp://phylomedb.org/phylomedb/proteomes/ ftp://phylomedb.org/phylomedb/proteomes/685557.1.fa.gz

noteworthy exception detecting 21 NRPSs in the genome of *Mortierella alpina*. The genes encoding typical PKS/fatty acid synthase (FAS) ketosynthase domains are most likely FAS alpha subunits: they reveal a characteristic domain order. BLAST searches using the tools at the specific home pages of the genomes confirm their annotation as FASs. Most of the discovered NRPS genes encode monomodular enzymes, except for the one in *Mortierella verticillata*, where we find a five-module protein encoded. Many NRPS-like genes do not possess the minimal set of domains necessary for the full enzymatic activity. These genes are therefore characterized as NRPS-like. Three genes which putatively encode for NRPSs were found in *Mucor circinelloides* f. *circinelloides* (Table 15.1) and can also be considered NRPS-like (Lee et al. 2014). Within the genus *Lichtheimia*, genes encoding PKSs, NRPSs, and DMATs are present in *L. hyalospora*, but absent in *L. corymbifera* (Schwartze et al. 2014). In some cases it can be supposed that the genes predicted actually represent the full enzymes but cannot be correctly annotated due to erroneous gene prediction. Another problem connected with genome assembly is the high AT content, which renders the bioinformation content low and prevents motif-based cluster prediction. Transcription regulators of secondary metabolism have not been yet systematically characterized, as reliable data is scarce. At this stage, we are mostly aware of pathway-specific regulators of clusters, but it is premature to draw general picture yet. In fact, known clusters in the *Ascomycetes* build the main basis of such genome-mining analyses, whereas information on proven clusters in other fungal phyla is lacking. In the *Ascomycetes*, about 60 % of PKS- and NRPS-encoding gene clusters include an embedded transcription regulator gene, which encodes in majority of cases a **zinc cluster transcription factor** (TF) (Brakhage 2013). The neighboring sequence regions of the genes encoding PKSs, NRPSs, and DMATs in zygomycetes were analyzed in order to confirm this preference for Zn cluster TFs. For this, we predicted that all TFs genome-wide extracted the TF annotations in regions of ±10 genes around those encoding secondary metabolite enzymes. We assigned these TFs to families of DNA-binding domains based on InterProScan predictions as described by Shelest (2008). Interestingly, Zn clusters are very modestly represented among these SM-accompanying TFs, leaving the first place to C_2H_2 Zn finger TFs and TFs of the homeodomain-like class. This observation becomes less surprising, however, if we think

about the overall predominance of C_2H_2 Zn fingers and especially of homeodomain-like TFs observed in zygomycetes. Our analysis suggests that every second, NRPS has a TF in the vicinity of 10 genes (8 of 15 NRPSs). For DMATs this number is higher (20 of 34, ~60 %). This corresponds to the number of the TFs in known ascomycete clusters (60 % for PKSs and NRPSs). The number for zygomycetes can be lower because we consider only the vicinity of 10 genes, whereas the cluster can be longer; on the other hand, considering longer region can give more false-positive predictions. The most frequent TFs in the vicinity of NRPSs and DMATs are homeodomain-like DNA-binding domain family TFs and C_2H_2 Zn finger TFs, comprising in sum nearly half of the total number of TFs that can potentially be the regulators of secondary metabolism in zygomycetes. This is not very surprising since these two families are the most abundant in at least *Mucorales* and *Entomophthorales* genomes (Schwartze et al. 2014). It is interesting to notice, however, that the most numerous families take over the regulation of the secondary metabolite clusters in fungi: in *Ascomycetes*, where Zn cluster is the dominating TF family, Zn cluster TFs are most frequently embedded in SM clusters, and in a similar picture we observe now for C_2H_2 and homeobox TFs in zygomycetes.

One promising strategy to explore and to broaden the biotechnological potential of the zygomycetes could be the investigation of zygomycetes in co-cultures with other microorganisms sharing the same ecological niche. This procedure has been shown successful in *Aspergillus* spp. for activation of silent gene clusters (Schroeckh et al. 2009; Nützmann et al. 2011, 2012) and is consistent with the cooperative nature of the zygomycetes as shown at the cellular and molecular level (Schachtschabel et al. 2008; Schimek and Wöstemeyer 2009; Krizsán et al. 2010; Voigt and Kirk 2014). Cocultivation of zygomycetes with other microorganisms sharing the same habitat under nature-close cultivation conditions has a high potential to increase the metabolic activity of zygomycetes, which are commonly known to be low producers of secondary compounds.

To sum it up, we show that all considered zygomycetes have a pronounced potential to produce natural products.

IV. Conclusions

- Zygomycetes have an integral role in the development of microbial ecosystems, a property which has the potential to be converted for biotechnological and industrial applications ranging from food technology to drug development.
- Understanding the biology, ecology, and biotrophic interactions of the zygomycetes with other microorganisms can help to explore novel secondary metabolites which are central to ecological functions and are useful for effective and innovative biotechnological utilizations,
- The genomic architecture and transcription factor repertoire of the zygomycetes largely differs from that of more recent fungal lineages. C_2H_2 transcription factors are predominant transcription factors.
- The degradation of β-carotene to pheromones has been extensively studied. The impact on biotechnological importance and applications has been neglected so far. Systematic–phylogenetic approaches may help with the screening for suitable production strains for biotechnological applications.

References

Abe A, Oda Y, Asano K, Sone T (2007) *Rhizopus delemar* is the proper name for Rhizopus oryzae fumaric-malic acid producers. Mycologia 99:714–722

Álvarez V, Rodríguez-Sáiz M, de la Fuente JL, Gudiña EJ, Godio RP, Martín JF, Barredo JL (2006) The crtS gene of *Xanthophyllomyces dendrorhous* encodes a novel cytochrome-P450 hydroxylase involved in the conversion of β-carotene into astaxanthin and other xanthophylls. Fungal Genet Biol 43:261–272

Anonymus (2007) Product focus: maleic anhydride. Chem Week 39

Appel KF, Wolff AM, Arnau J (2004) A multicopy vector system for genetic studies in Mucor circinel-

loides and other zygomycetes. Mol Genet Genomics 271:595–602
Arrach N, Fernández-Martín R, Cerdá-Olmedo E, Avalos J (2001) A single gene for lycopene cyclase, phytoene synthase, and regulation of carotene biosynthesis in *Phycomyces*. Proc Natl Acad Sci USA 98:1687–1692
Azeke MA, Greiner R, Jany KD (2011) Purification and characterization of two intracellular phytases from the tempeh fungus *Rhizopus oligosporus*. J Food Biochem 35:213–227
Bai DM, Li SZ, Liu ZL, Cui ZF (2008) Enhanced L-(+)-lactic acid production by an adapted strain of *Rhizopus oryzae* using corncob hydrolysate. Appl Biochem Biotechnol 144:79–85
Bartnicki-Garcia S (1968) Control of dimorphism in Mucor by hexoses: inhibition of hyphal morphogenesis. J Bacteriol 96:1586–1594
Beg QK, Kapoor M, Mahajan L, Hoondal GS (2001) Microbial xylanases and their industrial applications: a review. Appl Microbiol Biotechnol 56:326–338
Benny GL, Humber RA, Voigt K (2014) Zygomycetous Fungi: phylum Entomophthoromycota and subphyla Kickxellomycotina, Mortierellomycotina, Mucoromycotina, and Zoopagomycotina. In: McLaughlin DJ, Spatafora JW (eds) The Mycota, vol VIIA, 2nd edn, Systematics and evolution. Springer, Berlin, pp 209–250, Chapter 8
Bianciotto V, Lumini E, Bonfante P, Vandamme P (2003) "*Candidatus Glomeribacter gigasporarum*" gen. nov, sp. nov., an endosymbiont of arbuscular mycorrhizal fungi. Int J Syst Evol Microbiol 53:121–124
Bidartondo MI, Read DJ, Trappe JM, Merckx V, Ligrone R, Duckett JG (2011) The dawn of symbiosis between plants and fungi. Biol Lett 7:574–577
Bitar D, Van Cauteren D, Lanternier F, Dannaoui E, Che D, Dromer F, Desenclos JC, Lortholary O (2009) Increasing incidence of zygomycosis (mucormycosis), France, 1997-2006. Emerg Infect Dis 15:1395–1401
Bonito G, Gryganskyi A, Hameed K, Schadt C, Pelletier D, Schaefer A, Tuskan G, Labbe J, Martin F, Doktycz M, LaButti K, Ohm R, Grigoriev I, Vilgalys R (2013) Co-evolution of *Mortierella elongata* and its endosymbiotic bacterium. In: Abstracts submitted for presentation at the 2013 APS-MSA joint meeting, Austin, TX (2013-08-10–2013-08-14). St. Paul. American Phytopathological Society. Phytopathology 103 (6, Suppl 2):18–19 Presented at APS-MSA Joint Meeting
Brakhage AA (2013) Regulation of fungal secondary metabolism. Nat Rev Microbiol 11:21–32
Bulut S, Elibol M, Ozer D (2009) Optimization of process parameters and culture medium for L-(+)-lactic acid production by *Rhizopus oryzae*. J Chem Eng Jpn 42:589–595
Burmester A, Richter M, Schultze K, Voelz K, Schachtschabel D, Boland W, Wöstemeyer J, Schimek C (2007) Cleavage of β-carotene as the first step in sexual hormone synthesis in zygomycetes is mediated by a trisporic acid regulated β-carotene oxygenase. Fungal Genet Biol 44:1096–1108
Cao NJ, Du JX, Gong CS, Tsao GT (1996) Simultaneous production and recovery of fumaric acid from immobilized *Rhizopus oryzae* with a rotary biofilm contractor and adsorption column. Appl Environ Microbiol 62:2926–2931
Casadevall A, Pirofski L (2001) Host-pathogen interactions. The attributes of virulence. J Infect Dis 184 (3):337–344
Cavalier-Smith T (1981) Eukaryote kingdoms: seven or nine? BioSystems 14:461–481
Cavalier-Smith T (1998) A revised six-kingdom system of life. Biol Rev 73:203–266
Chakrabarti A, Chatterjee SS, Shivaprakash MR (2008) Overview of opportunistic fungal infections in India. Nihon Ishinkin Gakkai Zasshi 49:165–172
Chakrabarti A, Chatterjee SS, Das A, Panda N, Shivaprakash MR, Kaur A, Varma SC, Singhi S, Bhansali A, Sakhuja V (2009) Invasive zygomycosis in India: experience in a tertiary care hospital. Postgrad Med J 85:573–581
Chamilos G, Lewis RE, Kontoyiannis DP (2007) Multidrug-resistant endosymbiotic bacteria account for the emergence of zygomycosis: a hypothesis. Fungal Genet Biol 44(2):88–92
Chang Y, Wang S, Sekimoto S, Aerts AL, Choi C, Clum A, LaButti KM, Lindquist EA, Yee Ngan C, Ohm RA, Salamov AA, Grigoriev IV, Spatafora JW, Berbee ML (2015) Phylogenomic analyses indicate that early fungi evolved digesting cell walls of algal ancestors of land plants. Genome Biol Evol 7(6):1590–1601
Chatzifragkou A, Makri A, Belka A, Bellou S, Mavrou M, Mastoridou M, Mystrioti P, Onjaro G, Aggelis G, Papanikolaou S (2011) Biotechnological conversions of biodiesel derived waste glycerol by yeast and fungal species. Energy 36:1097–1108
Chaudhuri S, Ghosh S, Bhattacharyya DK, Bandyopadhyay S (1998) Effect of mustard meal on the production of arachidonic acid by *Mortierella elongata* SC-208. J Am Oil Chem Soc 75:1053–1055
Chayakulkeeree M, Ghannoum MA, Perfect JR (2006) Zygomycosis: the re-emerging fungal infection. Eur J Clin Microbiol Infect Dis 25:215–229
Csernetics Á, Nagy G, Iturriaga EA, Szekeres A, Eslava AP, Vágvölgyi CS, Papp T (2011) Expression of three isoprenoid biosynthesis genes and their effects on the carotenoid production of the zygomycete *Mucor circinelloides*. Fungal Genet Biol 48:696–703
Csernetics Á, Tóth E, Farkas A, Nagy G, Bencsik O, Vágvölgyi C, Papp T (2015) Expression of *Xanthophyllomyces dendrorhous* cytochrome-P450 hydroxylase and reductase in *Mucor circinelloides*. World J Microbiol Biotechnol 31:321–336

Czempinski K, Kruft V, Wöstemeyer J, Burmester A (1996) 4-dihydromethyltrisporate dehydrogenase from *Mucor mucedo*, an enzyme of the sexual hormone pathway:purification, and cloning of the corresponding gene. Microbiology 142:2647–2654

de Azevedo Santiago AL, de Souza Motta CM (2008) Isolation of Mucorales from processed maize (*Zea mays* L.) and screening for protease activity. Braz J Microbiol 39:698–700

de Hoog GS, Guarro J, Genè J, Figueras MJ (2000) Zygomycota. In: Atlas of clinical fungi, pp 58–124, Centraalbureau voor Schimmelcultures, Universitat Rovira I Virgili, Utrecht, The Netherlands

de Hoog GS, Ibrahim AS, Voigt K (2014) Emerging zygomycetes: an emerging problem in the clinical laboratory. Mycoses 57(Suppl 3):1

Dedyukhina E, Chistyakova T, Vainshtein M (2011) Biosynthesis of arachidonic acid by micromycetes (review). Appl Biochem Microbiol 47:109–117

Delalibera I Jr, Hajek AE, Humber RA (2003) Use of cell culture media for cultivation of the mite pathogenic fungi *Neozygites tanajoae* and *Neozygites floridana*. J Invertebr Pathol 84:119–127

Deng Y, Li S, Xu Q, Gao M, Huang H (2012) Production of fumaric acid by simultaneous saccharification and fermentation of starchy materials with 2-deoxyglucose-resistant mutant strains of *Rhizopus oryzae*. Bioresour Technol 107:363–367

Dolatabadi S, de Hoog GS, Meis JF, Walther G (2014a) Species boundaries and nomenclature of *Rhizopus arrhizus* (syn. *R. oryzae*). Mycoses 57:108–127

Dolatabadi S, Walther G, van den Ende AHGG, de Hoog GS (2014b) Diversity and delimitation of *Rhizopus microsporus*. Fungal Div 64:145–163

Dufossé L (2006) Microbial production of food grade pigments. Food Technol Biotechnol 44:313–321

Dufossé L (2008) Pigments from microalgae and microorganisms: sources of food colorants. In: Sociaciu C (ed) Food colorants, chemical and functional properties. CRC Press, Boca Raton, FL, pp 399–427

Dyal SD, Narine SS (2005) Implications for the use of *Mortierella* fungi in the industrial production of essential fatty acids. Food Res Int 38:445–467

Dyal SD, Bouzidi L, Narine SS (2005) Maximizing the production of γ-linolenic acid in *Mortierella ramanniana* var. *ramanniana* as a function of pH, temperature and carbon source, nitrogen source, metal ions and oil supplementation. Food Res Int 38:815–829

Fakas S, Papanikolaou S, Batsos A, Galiotou-Panayotou M, Mallouchos A, Aggelis G (2009) Evaluating renewable carbon sources as substrates for single cell oil production by *Cunninghamella echinulata* and *Mortierella isabellina*. Biomass Bioenergy 33:573–580

Ferreira JA, Lennartsson PR, Edebo L, Taherzadeh MJ (2013) Zygomycetes-based biorefinery: Present status and future prospects. Bioresour Technol 135:523–532

Fujimura R, Nishimura A, Ohshima S, Sato Y, Nishizawa T, Oshima K, Hattori M, Narisawa K, Ohta H (2014) Draft Genome Sequence of the Betaproteobacterial Endosymbiont Associated with the Fungus *Mortierella elongata* FMR23-6. Genome Announc 2(6):e01272-14

Gao D, Zeng J, Zheng Y, Yu X, Chen S (2013) Microbial lipid production from xylose by *Mortierella isabellina*. Bioresour Technol 133:315–321

Garre V, Barredo JL, Iturriaga EA (2015) Transformation of *Mucor circinelloides* f. *lusitanicus* protoplasts. In: van den Berg MA, Maruthachalam K (eds) Genetic transformation systems in fungi, Volume 1. Springer International Publishing, pp 49–59

Ghignone S, Salvioli A, Anca I, Lumini E, Ortu G, Petiti L, Cruveiller S, Bianciotto V, Piffanelli P, Lanfranco L, Bonfante P (2012) The genome of the obligate endobacterium of an AM fungus reveals an interphylum network of nutritional interactions. ISME J 6(1):136–145

Goldberg I, Rokem JS, Pines O (2006) Organic acids: old metabolites, new themes. J Chem Tech Biotechnol 81:1601–1611

Gooday GW (1974) Fungal sex hormones. Annu Rev Biochem 43:35–49

Gopinath SCB, Anbu P, Lakshmipriya T, Hilda A (2013) Strategies to characterize fungal lipases for applications in medicine and dairy industry. Biomed Res Int 2013:1545–1549

Greenberg RN, Scott LJ, Vaughn HH, Ribes JA (2004) Zygomycosis (mucormycosis): emerging clinical importance and new treatments. Curr Opin Infect Dis 17:517–525

Grundschober A, Tuor U, Aebi M (1998) *In vitro* cultivation and sporulation of *Neozygites parvispora* (Zygomycetes: Entomophthorales). Syst Appl Microbiol 21:461–469

Gryganskyi AP, Lee SC, Litvintseva AP, Smith ME, Bonito G, Porter T, Anishchenko IM, Heitman J, Vilgalys R (2010) Structure, function, and phylogeny of the mating locus in the *Rhizopus oryzae* complex. PLoS One 5(12), e15273

Guo Y, Yan Q, Jiang Z, Teng C, Wang X (2010) Efficient production of lactic acid from sucrose and corncob hydrolysate by a newly isolated Rhizopus oryzae GY18. J Ind Microbiol Biotechnol 37:1137–1143

Gutiérrez A, López-García S, Garre V (2011) High reliability transformation of the basal fungus *Mucor circinelloides* by electroporation. J Microbiol Met 84:442–446

Hawksworth DL (2011) A new dawn for the naming of fungi: impacts of decisions made in Melbourne in July 2011 on the future publication and regulation of fungal names. MycoKeys 1:7–20

Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL, Hedges SB (2001) Molecular evidence for

the early colonization of land by fungi and plants. Science 293:1129–1133
Hesseltine CW (1983) Microbiology of oriental fermented foods. Annu Rev Microbiol 37:575–601
Hesseltine CW (1991) Zygomycetes in food fermentations. Mycologist 5:162–169
Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson O, Huhndorf S, James T, Kirk PM, Lücking R, Lumbsch T, Lutzoni F, Matheny PB, McLaughlin DJ, Powell MJ, Redhead S, Schoch CL, Spatafora JW, Stalpers JA, Vilgalys R, Aime MC, Aptroot A, Bauer R, Begerow D, Benny GL, Castlebury LA, Crous PW, Dai Y-C, Gams W, Geiser DM, Griffith GW, Gueidan C, Hawksworth DL, Hestmark G, Hosaka K, Humber RA, Hyde K, Köljalb U, Kurtzman CP, Larsson K-H, Lichtwardt R, Longcore J, Miadlikowska J, Miller A, Moncalvo J-M, Mozley-Standridge S, Oberwinkler F, Parmasto R, Reeb V, Rogers JD, Roux C, Ryvarden L, Sampaio JP, Schuessler A, Sugiyama J, Thorn RG, Tibell L, Untereiner WA, Walker C, Wang A, Weir A, Weiss M, White M, Winka K, Yao Y-J, Zhang N (2007) A higher-level phylogenetic classification of the Fungi. Mycol Res 111:509–547
Hoffmann K, Telle S, Walther G, Eckart M, Kirchmair M, Prillinger HJ, Prazenica A, Newcombe G, Dölz F, Papp T, Vágvölgyi C, de Hoog S, Olsson L, Voigt K (2009) Diversity, genotypic identification, ultrastructural and phylogenetic characterization of zygomycetes from different ecological habitats and climatic regions: limitations and utility of nuclear ribosomal DNA barcode markers. In: Gherbawy Y, Mach RL, Rai M (eds) Current advances in molecular mycology. Nova, New York, pp 263–312
Hoffmann K, Voigt K, Kirk PM (2011) *Mortierellomycotina* subphyl. nov. based on multi-gene genealogies. Mycotaxon 115:353–363
Hoffmann K, Pawlowska J, Walther G, Wrzosek M, de Hoog GS, Benny GL, Kirk PM, Voigt K (2013) The family structure of the Mucorales: a synoptic revision based on comprehensive multigene-genealogies. Persoonia 30:57–76
Hou C (2008) Production of arachidonic acid and dihomo-linolenic acid from glycerol by oil-producing filamentous fungi, *Mortierella* in the ARS culture collection. J Ind Microbiol Biotechnol 35:501–506
Huang C, Chen XF, Xiong L, Chen XD, Ma LL, Chen Y (2013) Single cell oil production from low-cost substrates: the possibility and potential of its industrialization. Biotechnol Adv 31(2):129–139
Ibrahim AS (2011) Host cell invasion in mucormycosis: role of iron. Curr Opin Microbiol 14:406–411
Ibrahim AS, Gebremariam T, Liu M, Chamilos G, Kontoyiannis DP, Mink R, Kwon-Chung KJ, Fu Y, Skory CD, Edwards JE Jr, Spellberg B (2008) Bacterial Endosymbiosis is widely present among Zygomycetes but does not contribute to the pathogenesis of mucormycosis. J Infect Dis 198:1083–1090
Ikubo S, Takigawa N, Ueoka H, Kiura K, Tabata M, Shibayama T, Chikamori M, Aoe K, Matsushita A, Harada M (1999) *In vitro* evaluation of antimicrotubule agents in human small-cell lung cancer cell lines. Anticancer Res 19:3985–3988
Iturriaga EA, Díaz-Mínguez JM, Benito EP, Álvarez MI, Eslava AP (1992) Heterologous transformation of *Mucor circinelloides* with the *Phycomyces blakesleeanus leu1* gene. Curr Genet 21:215–223
Iturriaga EA, Velayos A, Eslava AP (2000) The structure and function of the genes involved in the biosynthesis of carotenoids in the Mucorales. Biotechnol Bioprocess Eng 5:263–274
James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung GH, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schüler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lücking R, Büdel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R (2006) Reconstructing the early evolution of the fungi using a six-gene phylogeny. Nature 443:818–822
Jang HD, Lin YY, Yang SS (2000) Polyunsaturated fatty acid production with *Mortierella alpina* by solid substrate fermentation. Bot Bull Acad Sin 41:41–48
Jennessen J, Nielsen KF, Houbraken J, Lyhne EK, Schnürer J, Frisvad JC, Samson RA (2005) Secondary metabolite and mycotoxin production by the *Rhizopus microsporus* group. J Agric Food Chem 53(5):1833–1840
Jin B, Yin P, Ma Y, Zhao L (2005) Production of lactic acid and fungal biomass by *Rhizopus* fungi from food processing waste water. J Ind Microbiol Biotechnol 32:678–686
Jones SK Jr, Bennett RJ (2011) Fungal mating pheromones: choreographing the dating game. Fungal Genet Biol 48:668–676
Karimi K, Zamani A (2013) *Mucor indicus*: biology and industrial application perspectives: a review. Biotechnol Adv 31:466–481
Kauffman CA (2004) Zygomycosis: reemergence of an old pathogen. Clin Infect Dis 39:588–590
Kavadia A, Komaitis M, Chevalot I, Blanchard F, Marc I, Aggelis G (2001) Lipid and gamma-linolenic acid accumulation in strains of zygomycetes growing on glucose. J Am Oil Chem Soc 78:341–346

Kloer DP, Ruch S, Al-Babili S, Beyer P, Schulz GE (2005) The structure of a retinal-forming carotenoid oxygenase. Science 8:267–269

Krisch J, Takó M, Papp T, Vágvölyi C (2010) Characteristics and potential use of β-glucosidases from Zygomycetes. In: Méndez-Vilas a (ed.) Current research, technology and education topics in applied microbiology and microbial biotechnology, Formatex Research Center, pp 891-896.

Krizsán K, Bencsik O, Nyilasi I, Galgóczy L, Vágvölgyi C, Papp T (2010) Effect of the sesterterpene-type metabolites, ophiobolins A and B, on zygomycetes fungi. FEMS Microbiol Lett 313:135–140

Lee S, Jang Y, Lee YM, Lee J, Lee H, Kim GH, Kim JJ (2011) Rice straw-decomposing fungi and their cellulolytic and xylanolytic enzymes. J Microbiol Biotechnol 21:1322–1329

Lee SC, Billmyre RB, Li A, Carson S, Sykes SM, Huh EY, Mieczkowski P, Ko DC, Cuomo CA, Heitman J (2014) Analysis of a food-borne fungal pathogen outbreak: virulence and genome of a *Mucor circinelloides* isolate from yogurt. mBio 5(4):e01390-14

Lennartsson PR, Karimi K, Edebo L, Taherzadeh MJ (2009) Effects of different growth forms of *Mucor indicus* on cultivation on dilute-acid lignocellulosic hydrolyzate, inhibitor tolerance, and cell wall composition. J Biotechnol 143:255–261

Lichtwardt RW, Manier JF (1978) Validation of the Harpellales and Asellariales. Mycotaxon 7:441–442

Liu XJ, Liu RS, Li HM, Tang YJ (2012) Lycopene production from synthetic medium by *Blakeslea trispora* NRRL2895(+) and 2896(-) in a stirred-tank fermenter. Bioprocess Biosyst Eng 35:739–749

Lukács GY, Papp T, Somogyvári F, Csernetics Á, Nyilasi I, Vágvölgyi CS (2009) Cloning of the *Rhizomucor miehei* 3-hydroxy-3-methylglutaryl-coenzyme A reductase gene and its heterologous expression in *Mucor circinelloides*. Ant Leeuwenhoek 95:55–64

Ma LJ, Ibrahim AS, Skory C, Grabherr MG, Burger G, Butler M, Elias M, Idnurm A, Lang BF, Sone T, Abe A, Calvo SE, Corrochano LM, Engels R, Fu J, Hansberg W, Kim JM, Kodira CD, Koehrsen MJ, Liu B, Miranda-Saavedra D, O'Leary S, Ortiz-Castellanos L, Poulter R, Rodriguez-Romero J, Ruiz-Herrera J, Shen YQ, Zeng Q, Galagan J, Birren BW, Cuomo CA, Wickes BL (2009) Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication. PLoS Genet 5, e1000549

Medina HR, Cerdá-Olmedo E, Al-Babili S (2011) Cleavage oxygenases for the biosynthesis of trisporoids and other apocarotenoids in *Phycomyces*. Mol Microbiol 82(1):199–208

Mehta BJ, Salgado LM, Bejarano ER, Cerdá-Olmedo E (1997) New mutants of *Phycomyces blakesleeanus* for beta-carotene production. Appl Environ Microbiol 63:3657–3661

Mendoza L, Vilela R, Voelz K, Ibrahim AS, Voigt K, Lee SC (2014) Human fungal pathogens of Mucorales and Entomophthorales, Chapter 27. In: Casadevall A, Mitchell AP, Berman J, Kwon-Chung KJ, Perfect JR, Heitman J (eds) Cold spring harbor perspectives: fungal pathogens. Cold Spring Harbor Press, Cold Spring Harb Perspect Med 5(4) pii: a019562

Meussen BJ, de Graaff LH, Sanders JP, Weusthuis RA (2012) Metabolic engineering of *Rhizopus oryzae* for the production of platform chemicals. Appl Microbiol Biotechnol 94:875–886

Millati R, Edebo L, Taherzadeh MJ (2005) Performance of *Rhizopus*, *Rhizomucor*, and *Mucor* in ethanol production from glucose, xylose, and wood hydrolyzates. Enzyme Microb Technol 36:294–300

Monteiro de Souza P, de Oliveira Magalhães P (2010) Application of microbial α-amylase in industry – a review. Braz J Microbiol 41(4):850–861

Morace G, Borghi E (2012) invasive mold infections: virulence and pathogenesis of mucorales. Int J Microbiol 2012, Article ID 349278, 5 p

Moss ST (1975) Commensalism of the trichomycetes. In: Batra LR (ed) Insect-fungus symbiosis: nutrition, mutualism and commensalism. Allanheld, Osmun & Co., Montclair, pp 175–227

Mullaney EJ, Ullah AH (2003) The term phytase comprises several different classes of enzymes. Biochem Biophys Res Commun 312:179–184

Mullaney EJ, Daly CB, Ullah AH (2000) Advances in phytase research. Adv Appl Microbiol 47:157–199

Münchberg U, Wagner L, Spielberg ET, Voigt K, Rösch P, Popp J (2012) Spatially resolved investigation of the oil composition in single intact hyphae of *Mortierella* spp. with micro-Raman spectroscopy. Biochim Biophys Acta 1831:341–349

Münchberg U, Wagner L, Rohrer C, Voigt K, Jahreis G, Rösch P, Popp J (2015) Quantitative assessment of the degree of lipid unsaturation in intact *Mortierella* by Raman microspectroscopy. Anal Bioanal Chem 407:3303–3311

Nagy LG, Petkovits T, Kovács GM, Voigt K, Vágvölgyi C, Papp T (2011) Where is the unseen fungal diversity hidden? A study of *Mortierella* reveals a large contribution of reference collections to the identification of fungal environmental sequences. New Phytol 191:789–794

Nagy G, Farkas A, Csernetics Á, Bencsik O, Szekeres A, Nyilasi I, Vágvölgyi CS, Papp T (2014) Transcription analysis of the three HMG-CoA reductase genes of *Mucor circinelloides*. BMC Microbiol 14:93

Nahas E (1988) Control of lipase production by *Rhizopus oligosporus* under various growth conditions. J Gen Microbiol 134:227–233

Navarro E, Lorca-Pascual JM, Quiles-Rosillo MD, Nicolas FE, Garre V, Torres-Martinez S, Ruiz-Vazquez RM (2001) A negative regulator of light-inducible carotenogenesis in *Mucor circinelloides*. Mol Gen Genomics 266:463–470

Navarro E, Peñaranda A, Hansberg W, Torres-Martínez S, Garre V (2013) A white collar 1-like protein

mediates opposite regulatory functions in *Mucor circinelloides*. Fungal Genet Biol 52:42–52
Nicolás-Molina FE, Navarro E, Ruiz-Vázquez RM (2008) Lycopene over-accumulation by disruption of the negative regulator gene crgA in *Mucor circinelloides*. Appl Microbiol Biotechnol 78:131–137
Nout MJR, Kiers JL (2005) Tempeh fermentation, innovation and functionality: update into the third millennium. J Appl Microbiol 98:789–805
Nützmann HW, Reyes-Dominguez Y, Scherlach K, Schroeckh V, Horn F, Gacek A, Schümann J, Hertweck C, Strauss J, Brakhage AA (2011) Bacteria-induced natural product formation in the fungus *Aspergillus nidulans* requires Saga/Ada-mediated histone acetylation. Proc Natl Acad Sci USA 108 (34):14282–14287
Nützmann HW, Schroeckh V, Brakhage AA (2012) Regulatory cross talk and microbial induction of fungal secondary metabolite gene clusters. Methods Enzymol 517:325–341
Obraztsova IN, Prados N, Holzmann K, Avalos J, Cerda-Olmedo E (2004) Genetic damage following introduction of DNA in *Phycomyces*. Fungal Genet Biol 41:68–180
Orlowsky M (1991) *Mucor* dimorphism. Microbiol Rev 55:234–258
Osmani SA, Scrutton MC (1985) The sub-cellular localisation and regulatory properties of pyruvate carboxylase from *Rhizopus arrhizus*. Eur J Biochem 147:119–128
Overman SA, Romano AH (1969) Pyruvate carboxylase of Rhizopus nigricans and its role in fumaric acid production. Biochem Biophys Res Commun 37:457–463
Papp T, Velayos A, Bartók T, Eslava AP, Vágvölgyi CS, Iturriaga EA (2006) Heterologous expression of astaxanthin biosynthesis genes in *Mucor circinelloides*. Appl Microbiol Biotechnol 69:526–531
Papp T, Csernetics Á, Nyilasi I, Ábrók M, Vágvölgyi CS (2010) Genetic transformation of Zygomycetes fungi. In: Rai MK, Kövics GJ (eds) Progress in mycology. Springer, Netherlands, pp 75–94
Papp T, Hoffmann K, Nyilasi I, Petkovits T, Wagner L, Vágvölgyi C, Voigt K (2011) *Mortierella*. In: Liu D (ed) Molecular detection of human fungal pathogens. Taylor & Francis CRC Press, pp 749–758
Papp T, Csernetics Á, Nagy G, Bencsik O, Iturriaga EA, Eslava AP, Vágvölgyi CS (2013) Canthaxanthin production with modified *Mucor circinelloides* strains. Appl Microbiol Biotechnol 97:4937–4950
Parniske M (2008) Arbuscular mycorrhiza: the mother of plant root endosymbioses. Nat Rev Microbiol 10:763–775
Partida-Martinez LP, Hertweck C (2005) Pathogenic fungus harbours endosymbiotic bacteria for toxin production. Nature 437:884–888
Partida-Martinez LP, Bandemer S, Rüchel R, Dannaoui E, Hertweck C (2008) Lack of evidence of endosymbiotic toxin-producing bacteria in clinical *Rhizopus* isolates. Mycoses 51(3):266–269
Petrič S, Hakki T, Bernhardt R, Zigon D, Crešnar B (2010) Discovery of a steroid 11α-hydroxylase from *Rhizopus oryzae* and its biotechnological application. J Biotechnol 150:428–437
Polaino S, Herrador MM, Cerdá-Olmedo E, Barrero AF (2010) Splitting of beta-carotene in the sexual interaction of Phycomyces. Org Biomol Chem 8 (19):4229–4231
Poliakov E, Gentleman S, Cunningham FX Jr, Miller-Ihli NJ, Redmond TM (2005) Key role of conserved histidines in recombinant mouse beta-carotene 15´15-monooxygenase-1 activity. J Biol Chem 280:29217–29223
Pritchard GG (1971) An NAD+-independent L-lactate dehydrogenase from *Rhizopus oryzae*. Biochim Biophys Acta 250:25–34
Pritchard GG (1973) Factors affecting the activity and synthesis of NAD dependent lactate dehydrogenase in *Rhizopus oryzae*. J Gen Microbiol 78:125–137
Quiles-Rosillo MD, Ruiz-Vázquez RM, Torres-Martínez S, Garre V (2003) Cloning, characterization and heterologous expression of the *Blakeslea trispora* gene encoding orotidine-5P-monophosphate decarboxylase. FEMS Microbiol Lett 222:229–236
Ratledge C (2004) Fatty acid biosynthesis in microorganisms being used for Single Cell Oil production. Biochimie 86:807–815
Ratledge C, Wynn JP (2002) The biochemistry and molecular biology of lipid accumulation in oleaginous microorganisms. Adv Appl Microbiol 51:1–51
Revuelta JL, Eslava AP (1983) A new gene (carC) involved in the regulation of carotenogenesis in *Phycomyces*. Mol Gen Genet 192:225–229
Ribes JA, Vanover-Sams CL, Baker DJ (2000) Zygomycetes in human disease. Clin Microbiol Rev 13 (2):236–301
Roa Engel CA, Straathof AJ, Zijlmans TW, van Gulik WM, van der Wielen LA (2008) Fumaric acid production by fermentation. Appl Microbiol Biotechnol 78:379–389
Rodríguez-Ortiz R, Michielse C, Rep M, Limón MC, Avalos J (2012) Genetic basis of carotenoid overproduction in *Fusarium oxysporum*. Fungal Genet Biol 49:684–696
Rodríguez-Sáiz M, Paz B, de la Fuente JL, López-Nieto MJ, Cabri W, Barredo JL (2004) *Blakeslea trispora* genes for carotene biosynthesis. Appl Environ Microbiol 70:5589–5594
Rodríquez-Sáiz M, de la Fuente JL, Barredo JL (2012) Metabolic engineering of *Mucor circinelloides* for zeaxanthin production. Methods Mol Biol 898:133–151
Roncero MIG, Cerdá-Olmedo E (1982) Genetics of carotene biosynthesis in *Phycomyces*. Curr Genet 5:5–8
Rothhardt J, Schwartze V, Voigt K (2011) Entomophthorales. In: Liu D (ed) Molecular detection

of human fungal pathogens. Taylor & Francis CRC Press, pp 723–734
Ruiz-Hidalgo MJ, Benito EP, Sandmann G, Eslava AP (1997) The phytoene dehydrogenase gene of *Phycomyces*: regulation of its expression by blue light and vitamin A. Mol Gen Genet 253:734–744
Ruiz-Hidalgo MJ, Eslava AP, Alvarez MI, Benito EP (1999) Heterologous expression of the *Phycomyces blakesleeanus* phytoene dehydrogenase gene (*carB*) in *Mucor circinelloides*. Curr Microbiol 39:259–264
Sahadevan Y, Richter-Fecken M, Kaerger K, Voigt K, Boland W (2013) Early and late trisporoids differentially regulate β-carotene production and gene transcript Levels in the mucoralean fungi *Blakeslea trispora* and *Mucor mucedo*. Appl Environ Microbiol 79(23):7466–7475
Saito K, Saito A, Ohnishi M, Oda Y (2004) Genetic diversity in *Rhizopus oryzae* strains as revealed by the sequence of lactate dehydrogenase genes. Arch Microbiol 182:30–36
Sajbidor J, Certik M, Dobronova S (1988) Influence of different carbon sources on growth, lipid content and fatty acid composition in four strains belonging to Mucorales. Biotechnol Lett 10:347–350
Salgado LM, Bejarano ER, Cerdá-Olmedo E (1989) Carotene superproducing mutants of *Phycomyces blakesleeanus*. Exp Mycol 13:332–336
Sanz C, Velayos A, Álvarez MI, Benito EP, Eslava AP (2011) Functional analysis of the *Phycomyces carRA* gene encoding the enzymes phytoene synthase and lycopene cyclase. PLos One 6:e23102
Sato Y, Narisawa K, Tsuruta K, Umezu M, Nishizawa T, Tanaka K, Yamaguchi K, Komatsuzaki M, Ohta H (2010) Detection of *Betaproteobacteria* inside the mycelium of the fungus *Mortierella elongata*. Microbes Environ 25(4):321–324
Sautour M, Soares Mansur C, Divies C, Bensoussan M, Dantigny P (2002) Comparison of the effects of temperature and water activity on growth rate of food spoilage moulds. J Ind Microbiol Biotechnol 28:311–315
Scarborough CL, Ferrari J, Godfray HC (2005) Aphid protected from pathogen by endosymbiont. Science 310(5755):1781
Schachtschabel D, Schimek C, Wöstemeyer J, Boland W (2005) Biological activity of trisporoids and trisporoid analogues in *Mucor mucedo* (-). Phytochemistry 66(11):1358–1365
Schachtschabel D, David A, Menzel KD, Schimek C, Wöstemeyer J, Boland W (2008) Cooperative biosynthesis of trisporoids by the (+) and (-) mating types of the zygomycete *Blakeslea trispora*. ChemBioChem 9(18):3004–3012
Schimek C, Wöstemeyer J (2009) Carotene derivatives in sexual communication of zygomycete fungi. Phytochemistry 70:1867–1875
Schimek C, Kleppe K, Saleem AR, Voigt K, Burmester A, Wöstemeyer J (2003) Sexual reactions in Mortierellales are mediated by the trisporic acid system. Mycol Res 107:736–747
Schroeckh V, Scherlach K, Nützmann HW, Shelest E, Schmidt-Heck W, Schuemann J, Martin K, Hertweck C, Brakhage AA (2009) Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in *Aspergillus nidulans*. Proc Natl Acad Sci USA 106(34):14558–14563
Schwartz SH, Tan BC, Gage DA, Zeevaart JA, McCarty DR (1997) Specific oxidative cleavage of carotenoids by VP14 of maize. Science 276:1872
Schwartze VU, Winter S, Shelest E, Marcet-Houben M, Horn F, Wehner S, Linde J, Valiante V, Sammeth M, Riege K, Nowrousian M, Kaerger K, Jacobsen ID, Marz M, Brakhage AA, Gabaldón T, Böcker S, Voigt K (2014) Gene expansion shapes genome architecture in the human pathogen *Lichtheimia corymbifera*: an evolutionary genomics analysis in the ancient terrestrial Mucorales (Mucoromycotina). PLoS Genet 10(8):e1004496. doi:10.1371/journal.pgen.1004496
Serrano I, Lopes da Silva T, Carlos Roseiro J (2001) Ethanol-induced dimorphism and lipid composition changes in Mucor fragilis CCMI 142. Lett Appl Microbiol 33:89–93
Sharifia M, Karimi K, Taherzadeh MJ (2008) Production of ethanol by filamentous and yeast-like forms of Mucor indicus from fructose, glucose, sucrose, and molasses. J Ind Microbiol Biotechnol 35:1253–1259
Shelest E (2008) Transcription factors in fungi. FEMS Microbiol Lett 286:145–151
Shelest E, Voigt K (2014) Genomics to study basal lineage fungal biology: phylogenomics suggests a common origin. In: Nowrousian M (ed) The Mycota, vol XIII, 2nd edn, Fungal genomics. Springer, Berlin, pp 31–60
Sigel R, Sigel A, Sigel H (2007) The ubiquitous roles of cytochrome P450 proteins: metal ions in life sciences. Wiley, New York. ISBN 0-470-01672-8
Silva F, Torres-Martinez S, Garre V (2006) Distinct white collar-1 genes control specific light responses in *Mucor circinelloides*. Mol Microbiol 61:1023–1037
Silva F, Navarro E, Peñaranda A, Murcia-Flores L, Torres-Martínez S, Garre V (2008) A RING-finger protein regulates carotenogenesis via proteolysis-independent ubiquitylation of a white collar-1-like activator. Mol Microbiol 70:1026–1036
Simon L, Bousquet J, Lévesque RC, Lalonde M (1993) Origin and diversification of endomycorrhizal fungi and co-incidence with vascular land plants. Nature 363:67–69
Skory CD (2000) Isolation and expression of lactate dehydrogenase genes from *Rhizopus oryzae*. Appl Environ Microbiol 66:2343–2348
Spano LA, Medeiros J, Mandels M (1976) Enzymatic hydrolysis of cellulose wastes to glucose. Resour Recover Conserv 1:279–294

Stredansky M, Conti E, Stredanska S, Zanetti F (2000) γ-Linolenic acid production with *Thamnidium elegans* by solid-state fermentation on apple pomace. Bioresour Technol 73:41–45

Sun J, Sun XX, Tang PW, Yuan QP (2012) Molecular cloning and functional expression of two key carotene synthetic genes derived from *Blakeslea trispora* into *E. coli* for increased β-carotene production. Biotechnol Lett 34:2077–2082

Tagua VG, Medina HR, Martín-Domínguez R, Eslava AP, Corrochano LM, Cerdá-Olmedo E, Idnurm A (2012) A gene for carotene cleavage required for pheromone biosynthesis and carotene regulation in the fungus *Phycomyces blakesleeanus*. Fungal Genet Biol 49(5):398–404

Takahashi Y, Moiseyev G, Chen Y, Ma JX (2005) Identification of conserved histidines and glutamic acid as key residues for isomerohydrolase activity of RPE65, an enzyme of the visual cycle in the retinal pigment epithelium. FEBS Lett 579:5414–5418

Torres-Martínez S, Murillo FJ, Cerdá-Olmedo E (1980) Genetics of locopene cyclization and substrate transfer in beta-carotene biosynthesis in *Phycomyces*. Genet Res 36:299–946

Tsuruo T, Oh-hara T, Iida H, Tsukagoshi S, Sato Z, Matsuda I, Iwasaki S, Okuda S, Shimizu F, Sasagawa K, Fukami M, Fukada K, Arakawa M (1986) Rhizoxin, a macrocyclic lactone antibiotic, as a new antitumor agent against human and murine tumor cells and their vincristine-resistant sublines. Cancer Res 46:381–385

van der Westhuizen JPJ, Kock JLF, Botha A, Botes PJ (1994) The distribution of the ω3- and ω 6-series of cellular long-chain fatty acids in fungi. Sys Appl Microbiol 17:327–345

van Heeswijck R, Roncero MIG (1984) High frequency transformation of *Mucor* with recombinant plasmid DNA. Carlsberg Res Commun 49:691–702

Velayos A, Blasco JL, Alvarez MI, Iturriaga EA, Eslava AP (2000a) Blue-light regulation of the phytoene dehydrogenase (carB) gene expression in Mucor circinelloides. Planta 210:938–946

Velayos A, Eslava AP, Iturriaga EA (2000b) A bifunctional enzyme with lycopene cyclase and phytoene synthase activities is encoded by the carRP gene of Mucor circinelloides. Eur J Biochem 267:1–12

Velayos A, Papp T, Aguilar-Elena R, Fuentes-Vicente M, Eslava AP, Iturriaga EA, Álvarez MI (2003) Expression of the *carG* gene, encoding geranylgeranyl pyrophosphate synthase, is up-regulated by blue light in *Mucor circinelloides*. Curr Genet 43:112–120

Velayos A, Fuentes-Vicente M, Aguilar-Elena R, Eslava AP, Iturriaga EA (2004) A novel fungal prenyl diphosphate synthase in the dimorphic zygomycete *Mucor circinelloides*. Curr Genet 45:371–377

Vially G, Marchal R, Guilbert N (2010) L(+) Lactate production from carbohydrates and lignocellulosic materials by *Rhizopus oryzae* UMIP 4.77. World J Microb Biot 26:607–614

Voigt K (2012) Zygomycota. In: Frey W (ed) Syllabus of plant families – A. Engler's Syllabus der Pflanzenfamilien. Part 1/1: Blue-green algae, Myxomycetes and Myxomycete-like organisms, Phytoparasitic protists, Heterotrophic Heterokontobionta and Fungi p.p. Borntraeger Verlag, Stuttgart, pp 130–162

Voigt K, de Hoog GS (2013) The zygomycetes in a phylogenetic perspective. Special issue Persoonia: molecular phylogeny and evolution of fungi, vol. 30. Naturalis Biodiversity Center Leiden and Centraalbureau voor Schimmelcultures Utrecht, The Netherlands, p 125

Voigt K, Kirk PM (2014) 136. FUNGI | Classification of the zygomycetes: reappraisal as coherent class based on a comparison between traditional *versus* molecular systematics. In: Batt CA, Tortorello ML (eds) Encyclopedia of food microbiology, Vol 2. Elsevier, Academic, pp 54–67

Voigt K, Wöstemeyer J (2001) Phylogeny and origin of 82 zygomycetes from all 54 genera of the Mucorales and Mortierellales based on combined analysis of actin and translation elongation factor EF-1α genes. Gene 270:113–120

Voigt K, Hoffmann K, Einax E, Eckart M, Papp T, Vágvölgyi C, Olsson L (2009) Revision of the family structure of the Mucorales (Mucoromycotina, Zygomycetes) based on multigene-genealogies: phylogenetic analyses suggest a bigeneric Phycomycetaceae with *Spinellus* as sister group to *Phycomyces*. In: Gherbawy Y, Mach RL, Rai M (eds) Current advances in molecular mycology. Nova, New York, pp 313–332

Voigt K, Marano AV, Gleason F (2013) Ecological and economical importance of parasitic and zoosporic true fungi. In: Kempken F (ed) The Mycota, vol XI, 2nd edn, Agricultural applications. Springer, Berlin, pp 243–270

Wagner L, Stielow B, Hoffmann K, Petkovits T, Papp T, Vágvölgyi C, de Hoog GS, Verkley G, Voigt K (2013) A comprehensive molecular phylogeny of the Mortierellales (Mortierellomycotina) based on nuclear ribosomal DNA. Persoonia 30:77–93

Wang GY, Keasling JD (2002) Amplification of HMG-CoA reductase production enhances carotenoid accumulation in *Neurospora crassa*. Metab Eng 4:193–201

Wang L, Chen W, Feng Y, Ren Y, Gu Z et al (2011) Genome characterization of the oleaginous fungus *Mortierella alpina*. PLoS One 6:e28319

Wang D, Wu R, Xu Y, Li M (2013) Draft genome sequence of *Rhizopus chinensis* CCTCCM201021, used for brewing traditional Chinese alcoholic beverages. Genome Announc 1(2), e0019512

Werpy T, Petersen G (2004) Top value added chemicals from biomass, vol 1, Results of screening for potential candidates from sugars and synthesis gas. US Department of Energy, Washington, DC

Wetzel J, Scheibner O, Burmester A, Schimek C, Wöstemeyer J (2009) 4-Dihydrotrisporin-dehydrogenase, an enzyme of the sex hormone pathway of *Mucor mucedo*: purification, cloning of the corresponding gene, and developmental expression. Eukaryotic Cell 8:88–95

White JD, Blakemore PR, Green NJ et al (2002) Total synthesis of rhizoxin D, a potent antimitotic agent from the fungus *Rhizopus chinensis*. J Org Chem 67:7750–7760

Whittaker RH (1969) New concepts of kingdoms of organisms. Science 163:150–160

Wikandari R, Millati R, Lennartsson PR, Harmayani E, Taherzadeh MJ (2012) Isolation and characterization of zygomycetes fungi from tempeh for ethanol production and biomass applications. Appl Biochem Biotechnol 167:1501–1512

Willke T, Vorlop KD (2004) Industrial bioconversion of renewable resources as an alternative to conventional chemistry. Appl Microbiol Biotechnol 66:131–142

Wolff AM, Arnau J (2002) Cloning of glyceraldehyde-3-phosphate dehydrogenase-encoding genes in *Mucor circinelloides* (Syn. *racemosus*) and use of the *gpd1* promoter for recombinant protein production. Fungal Genet Biol 35:21–29

Wöstemeyer J, Burmester A, Wöstemeyer A, Schultze K, Voigt K (2002) Gene transfer in the fungal host-parasite system *Absidia glauca-Parasitella parasitica* depends on infection. In: Syvanen M, Kado CI (eds) Horizontal gene transfer, 2nd edn. Academic, San Diego, CA, pp 241–247, Chapter 21

Wöstemeyer J, Grünler A, Schimek C, Voigt K (2005) Genetic regulation of carotenoid biosynthesis in fungi. In: Applied mycology and biotechnology Vol 5: Genes and genomics. Elsevier, pp 257–274

Yen HW, Lee YC (2010) Production of lactic acid from raw sweet potato powders by rhizopus oryzae immobilized in sodium alginate capsules. Appl Biochem Biotechnol 162:607–615

Zeng J, Zheng Y, Yu X, Yu L, Gao D, Chen S (2013) Lignocellulosic biomass as a carbohydrate source for lipid production by *Mortierella isabellina*. Bioresour Technol 128:385–391

Zhang ZY, Jin B, Kelly JM (2007) Production of lactic acid from renewable materials by *Rhizopus* fungi. Biochem Eng J 35:251–263

Zorn H, Langhoff S, Scheibner M, Berger RG (2003) Cleavage of β-carotene to flavor compounds by fungi. Appl Microbiol Biotechnol 62(4):331–336

3 Diskussion

3.1 Gründe für die Existenz von Sekundärmetabolit-Gen-Clustern

Sekundärmetabolit-Gen-Cluster (SMGCs) sind das Vorhersageziel der im Zuge dieser Arbeit entwickelten Methoden. Daher stellt sich die grundlegende Frage, warum bestimmte Gene überhaupt in Clustern organisiert sind. Die vorliegende Arbeit kann diese Frage nicht beantworten. Dennoch sollen im Folgenden die gängigsten Hypothesen zur Existenz von SMGCs kurz diskutiert werden. Sicher ist, dass sowohl in Prokaryoten als auch in Eukaryoten die Gene nicht zufällig im Genom verteilt sind [Lawrence 1999; Hurst u. a. 2004].

Eine der ersten Erklärungen für die Existenz von SMGCs in eukaryotischen Genomen war der horizontale Gentransfer[1]. Zum Beispiel könnten Pilze den Penizillinstoffwechselweg als Teil des Erbgutes von Bakterien erhalten haben [Keller und Hohn 1997]. Dies allein erklärt aber nicht, warum sich die Gene in den Pilzgenomen im Laufe der Evolution nicht wieder voneinander entfernt haben und warum sie zuvor bereits im Bakteriengenom in einem Cluster organisiert waren. Do und Miyano [2008] stellten fest, dass die Herkunft der SMGCs in *Aspergillus fumigatus* nicht allein durch horizontalen Gentransfer erklärt werden kann.

Laut Yi u. a. [2007] ist die koordinierte Expression einer Gruppe von Genen, als Voraussetzung für eine effiziente Biosynthese, leichter in einem eng gekoppelten Cluster zu erreichen als wenn die entsprechenden Gene zufällig im Genom verteilt wären. Dadurch könnte ein Selektionsdruck entstehen, der die Bildung und Erhaltung von Gen-Clustern fördert.

[1] Als *horizontaler Gentransfer* wird die Übertragung von Erbmaterial zwischen zwei Organismen unterschiedlicher Abstammungslinien bezeichnet. Auf diese Weise können zum Beispiel Antibiotikaresistenzen zwischen verschiedenen Bakterienarten übertragen werden.

McGary u. a. [2013] argumentieren, dass die Verweildauer von meist giftigen Zwischenprodukten der Sekundärmetabolitsynthese in der Zelle durch Gen-Cluster verringert werden kann. Sie fanden heraus, dass in Pilzgenomen häufig zwei Gene benachbart sind, die mit dem selben toxischen Metaboliten hinsichtlich Aufbau, Abbau oder Umwandlung in Verbindung stehen. Zusätzlich spielt die Orientierung der Paare eine wichtige Rolle: Umso toxischer das Zwischenprodukt, desto öfter sind Gene entgegengesetzt orientiert. Das heißt, sie werden in den meisten Fällen von einem gemeinsamen bidirektionalen Promotor reguliert. Jedoch beschränkte sich diese Arbeit auf die Untersuchung von Genpaaren (»kleinstmögliche Cluster«). Die Vermeidung von giftigen Zwischenprodukten erklärt damit nur das häufige Vorkommen solcher Paare, nicht aber die Entstehung von Gen-Clustern mit drei und mehr Genen.

In Eukaryoten sind viele Beispiele von »lose« geclusterten Genen bekannt [Lee und Sonnhammer 2003; Walker u. a. 2012; Ghanbarian und Hurst 2015]. Diese sind meist nicht direkt benachbart, sondern über einen größeren Genomabschnitt verteilt. Auch ein funktioneller Zusammenhang ist nicht immer gegeben. Ein Vergleich mit SMGCs ist daher kaum möglich. Interessant ist, dass zum Beispiel bei Säugetieren und bei *Saccharomyces cerevisiae* (Bäckerhefe) essentielle Gene häufig geclustert sind [Rubin und Green 2013]. Im Gegensatz dazu sind die nichtessentiellen Gene des Sekundärmetabolismus bei Pilzen häufig geclustert (Abschnitt 1.2).

Lemay u. a. [2012] fassten zusammen, dass eine Mischung aus den folgenden Mechanismen verantwortlich ist für die Entstehung und Erhaltung von Gen-Clustern: Genduplikation, epigenetische Regulation, Unterdrückung von Stop-Codons[2], gemeinsame Transkriptionsfaktorbindestellen, ähnliche Funktionsweise der kodierten Proteine, aufeinander aufbauende Funktion der kodierten Proteine und Gewebespezitifität. Aufgrund von verschiedenen Cluster-Definitionen, unterschiedlichen Genomstrukturen und teilweise widersprüchlichen Erkenntnissen ist ein einheitliches und klares Verständnis über Gen-Cluster im Allgemeinen und SMGCs im Speziellen nach wie vor nicht gegeben [Michalak 2008].

Schließlich bleibt die Frage, warum andere Gene gerade nicht in Clustern organisiert sind. Scheinbar profitiert nur ein Teil der Gene eines Genoms von dieser Strukturie-

[2]Beim *»translational readthrough«* wird durch eine Mutation ein Stop-Codon nicht als solches interpretiert. Dadurch wird die Translation beim nächsten Gen ohne Unterbrechung fortgesetzt, wenn dieses in die gleiche Richtung zeigt.

rung. Für den anderen Teil scheint der Verlust an Freiheit bei der Genomorganisation keinen ausreichend hohen evolutionären Vorteil zu bieten oder sogar von Nachteil zu sein [Liao und Zhang 2008].

3.2 SMIPS

Wie in Manuskript 3 (Abschnitt 2.4) beschrieben, ist SMIPS ein Programm zur Vorhersage von Ankergenen in einer Menge von Proteinsequenzen. SMIPS wurde zum einen entwickelt, um auf unkomplizierte Weise und unabhängig von anderen Programmen, zum Beispiel SMURF, Ankergene vorhersagen zu können. Zum anderen kann die Ausgabe von SMIPS, eine Liste von Ankergenen, als Eingabe für CASSIS genutzt werden. Eine weitere Besonderheit von SMIPS ist, dass es die Abfolge und Annotation der Proteindomänen in den Ankergenen in einer für Menschen leicht verständlichen und gleichzeitig für Computer leicht verarbeitbaren Weise darstellt.

Im Gegensatz zu anderen Programmen zur Ankergenvorhersage, die meist auf spezielle Enzymklassen beschränkt sind (Abschnitt 1.4, Weber [2014]) kann SMIPS universell für die Vorhersage von PKSs Typ I, II und III, NRPSs, DMATSs, und allen möglichen PKS/NRPS/DMATS-Hybriden genutzt werden. SMIPS erkennt Ankergene, indem es die Proteinsequenzen auf Domänen hin untersucht, die typischerweise in Ankergenen vorkommen. In diesem Punkt unterscheidet sich SMIPS kaum von vergleichbaren Programmen. Der Unterschied ist allerdings, dass es die äußerst umfangreiche Datenbank InterPro[3] nutzt, welche wiederum auf die Einträge von elf verschiedenen Datenbanken mit Proteinannotationen zurückgreift. SMURF zum Beispiel nutzt zur Vorhersage von Ankergenen nur die beiden Datenbanken Pfam und TIGRPFAM [Khaldi u. a. 2010].

Ein Nachteil von SMIPS ist, dass die InterPro IDs (IPRs) typischer Ankergene manuell gesammelt werden müssen. Die von SMIPS genutzte Liste von IPRs repräsentiert typische Ankergendomänen von Pilzen, Bakterien und Pflanzen (Unterabschnitt 2.4.1, Tabelle S2). Jedoch stammt ein Großteil der IPRs, aufgrund der ungleichmäßigen Verteilung der verfügbaren Daten, von Ankergenen verschiedener *Aspergillus*-Arten. Unabhängig davon kann SMIPS auf jedes beliebige Genom zur Ankergenvorhersage angewandt werden, liefert für Pilze aber vermutlich die genauesten Ergebnisse.

[3] *InterPro* [Jones u. a. 2014] ist eine »Meta-Datenbank« zur funktionellen Analyse und Klassifikation von Proteinfamilien, Proteindomänen und aktiven Zentren.

3.3 Anwendung der *de novo* Motivsuche in CASSIS

3.3.1 Anzahl und Länge der Promotorsequenzen für die Motivsuche

CASSIS benötigt als Eingabe eine Genomsequenz und die dazugehörige Position eines jeden Gens im Genom. Mit Hilfe dieser Informationen wird für jedes Gen eine Promotorsequenz ermittelt. Für die *de novo* Motivsuche (Unterabschnitt 1.6.1) übergibt CASSIS maximal 15 Promotoren vor und nach einem gegebenen Ankergen an MEME. Die Frage ist nun, warum nicht mehr und/oder längere Sequenzen, zum Beispiel die gesamte intergenische Region, übergeben werden.

Zum einen beruht CASSIS maßgeblich auf der Hypothese der Kolokalisation und Koregulation der Cluster-Gene (Abschnitt 1.2). Das heißt, die Bindestellen eines cluster-spezifischen Transkriptionsfaktors sind in den Cluster-Genen angereichert und kommen außerhalb des Clusters nur selten vor, unter anderem durch falsch positive Motivfunde beim »Scannen« (Unterabschnitt 1.6.1) oder einfach zufällig[4]. Ein Bereich von 30 Promotoren beziehungsweise Genen (Position des Ankergens ±15 Promotoren) ist etwas größer als die größten bekannten SMGCs (zum Beispiel Aflatoxin, 25 Gene, Yu u. a. [2004]). Daher wurde auf die Einbeziehung von weiteren Promotoren bei der *de novo* Motivsuche verzichtet. Tests ergaben, dass weitere Promotoren das Ergebnis der Motivsuche nicht verbessern. Im Gegenteil, mehr Eingabesequenzen, die das Bindestellenmotiv (mutmaßlich) nicht enthalten, erhöhen die Laufzeit der *de novo* Motivsuche und die Wahrscheinlichkeit falsch positiv vorhergesagter Motive [Hu u. a. 2005].

Das Gleiche trifft auf unnötig lange Promotorsequenzen zu. Wie in Manuskript 3 (Abschnitt 2.4) beschrieben, wurden die Einträge der Datenbanken TRANSFAC [Matys u. a. 2003] und FunTF[5] hinsichtlich der Position aller bekannten Transkriptionsfaktoren bei Pilzen untersucht. Alle Einträge dieser Datenbanken sind mit Publikationen belegt. Ein Großteil der Bindestellen befindet sich in einem Bereich von 1000 bp vor und 50 bp nach der Transkriptionsstartstelle (TSS) des regulierten Gens. Daher wird nur dieser Sequenzabschnitt für die Motivsuche (*de novo* und

[4] Sequenzmotive von TFBSs sind, im Vergleich zur Gesamtlänge des Genoms, sehr kurz. Daher besteht die Möglichkeit, dass zufällig andere Stellen im Genom die gleiche oder eine ähnlich Abfolge von Nukleotiden aufweisen.

[5] https://sbi.hki-jena.de/funtf, August 2015

»scannen«) genutzt, und nicht zum Beispiel die gesamte Sequenz zwischen zwei Genen. Bei intergenischen Regionen kürzer als 1000 bp wird die Promotorsequenz automatisch verkürzt. Bei zwei bidirektionalen[6] Genen, mit einem Abstand von 2000 bp und weniger, werden deren Promotoren zu einem bidirektionalen Promotor zusammengefasst.

3.3.2 Warum ausschließlich MEME für die Motivsuche genutzt wird

Im folgenden Abschnitt wird erklärt, warum bei der Entwicklung von CASSIS ausschließlich MEME für die *de novo* Motivsuche zum Einsatz kommt und keine der anderen in Abschnitt 1.6 vorgestellten Methoden.

Um Motivsuchealgorithmen zu vergleichen, wird eine *konstante Menge* an Eingabesequenzen ausgewählt oder künstlich generiert, und verschiedene Programme auf diese angewandt. Danach werden die Ergebnisse verglichen und eine Aussage getroffen, wie »gut« die einzelnen Algorithmen abgeschnitten haben [Tompa u. a. 2005; Sandve u. a. 2007]. Im Gegensatz zu diesem Testverfahren erzeugt CASSIS bis zu 250 *verschiedene Mengen* an Eingabesequenzen, die dann in verschiedenen Motiven resultieren. Daher spielt für CASSIS die Auswahl der Motive eine wichtigere Rolle als die Auswahl des Motivsuchealgorithmus. Die Auswahl der »richtigen« Motive war einer der Schwerpunkte bei der Entwicklung von CASSIS und ist in Manuskript 3 im Detail beschrieben. Dennoch bleibt die Frage bestehen, warum MEME für die *de novo* Motivsuche genutzt wurde. Dies hat verschiedene Gründe:

Für MEME ist nicht nur eine grafische Oberfläche, sondern auch eine Kommandozeilenversion verfügbar. Ohne diese wäre die Integration in CASSIS nicht möglich gewesen.

MEME erlaubt pro Promotor null bis beliebig viele Vorkommen des vorherzusagenden Motivs (»ANR«[7]). Viele andere Programme unterstützen nur das »ZOOPS-Modell[8]« [Pavesi u. a. 2004]. ANR ist wichtig, da ein Promotor mehrere (funktionale)

[6] Beide Gene liegen auf verschiedenen DNA-Strängen und die Richtung ihrer Transkription ist voneinander weg orientiert.

[7] Bei *»any number of repetition« (ANR)* kann jede Eingabesequenz der Motivsuche null bis beliebig viele Vorkommen des vorherzusagenden Motivs enthalten.

[8] Bei *»zero or one occurrence per sequence« (ZOOPS)* wird angenommen, dass jede Eingabesequenz das vorherzusagende Motiv gar nicht oder genau einmal enthält.

Bindestellen für den cluster-spezifischen Transkriptionsfaktor enthalten kann [Ehrlich u. a. 1999, 2002]. Zusätzlich kann es »Lückengene« mit gar keiner Bindestelle geben [Schroeckh u. a. 2009]. Schließlich werden im Zuge der Ermittlung der Promotorsequenzen (Unterabschnitt 3.3.1) oftmals bidirektionale Promotoren erzeugt, die wahrscheinlich mindestens zwei Bindestellen enthalten.

MEME stellt keine besonderen Anforderungen an die Eingabesequenzen, wie zum Beispiel, dass sie gleich lang sein müssen. Diese Eigenschaft ist ebenfalls wichtig, da die von CASSIS erzeugten Promotorsequenzen 6 bp (minimale Motivlänge) bis 2101 bp (bidirektionaler Promotor[9]) lang sein können. Außerdem bezieht MEME selbstständig den Gegenstrang (reverses Komplement) aller Promotoren in die Motivsuche mit ein.

Die Länge der zu suchenden Sequenzmotive ist im Allgemeinen unbekannt. Laut Fazius u. a. [2011] sind TFBSs 6–12 bp lang. Diese minimale und maximale Länge wird bei jedem MEME-Aufruf als Parameter angegeben. Viel wichtiger ist aber, dass MEME die tatsächliche Motivlänge innerhalb dieses Bereiches selbstständig in einem Durchlauf anpassen kann und dadurch im Vorteil ist gegenüber Algorithmen mit fester Motivlänge [Hu u. a. 2005].

Schließlich können die von MEME ausgegebenen Motive (Profile), ohne Umwandlungszwischenschritte, als Eingabe für FIMO [Grant u. a. 2011] genutzt werden. Andere Programme erfordern die Umwandlung in ein anderes Format oder in einzelne Sequenzen. Mit FIMO wird später das Genom nach allen Vorkommen des von MEME vorhergesagten Motivs durchsucht (»gescannt«, Unterabschnitt 1.6.1).

3.3.3 Anzahl und Überprüfung der von MEME gefundenen Motive

Die meisten Programme zur *de novo* Motivsuche sind in der Lage, verschiedene Motive in den Eingabesequenzen zu erkennen. MEME zum Beispiel kann nach genau so vielen verschiedenen Motiven suchen, wie beim Programmaufruf angegeben werden. Nachdem MEME ein Motiv gefunden hat, markiert es die Vorkommen dieses Motivs in den Eingabesequenzen und beginnt nach dem nächsten Motiv zu suchen [Bailey und Elkan 1993]. In der Ausgabe von MEME werden die verschiedenen Motive nach

[9] $2101 = (1000 + 50 + 1\ \text{[TSS]}\ ) \cdot 2 - 1$ [minimale Überschneidung von 1 bp]

ihrem E-Wert[10] sortiert. MEME gibt das »beste« Motiv immer zuerst aus.

Laut Hu u. a. [2005] stellt das von verschiedenen Motivsuchealgorithmen ermittelte »beste« Motiv oft nicht die tatsächlich beste Vorhersage dar. Dennoch lässt CASSIS von MEME nur ein Motiv suchen – warum? Zum einen würde die Berechnung von zum Beispiel zwei Motiven die Laufzeit der bis zu 250 MEME-Aufrufe und alle Folgeschritte von CASSIS ungefähr verdoppeln. Zum anderen wird ein Motiv, dass an zweiter Stelle erscheint, mit hoher Wahrscheinlichkeit bei mindestens einer der anderen 249 Eingaben an erster Stelle erscheinen. Falls nicht, dann ist davon auszugehen, dass dieses Motiv kein Kandidat für eine cluster-spezifische TFBS ist. Daher ist die Suche nach mehr Motiven nicht unbedingt notwendig.

Experimentell ermittelte Bindestellensequenzen sind aktuell nur für wenige cluster-spezifische Transkriptionsfaktoren verfügbar. In Bezug auf die in dieser Arbeit verwendeten 38 bekannten SMGCs sind das die Bindestellen für die Transkriptionsfaktoren AflR aus dem Aflatoxin-Gen-Cluster (Ehrlich u. a. [1999], Abbildung 3.1) und GliZ aus dem Gliotoxin-Gen-Cluster (Daniel Scharf, persönliche Mitteilung, Abbildung 3.2). Beide besitzen eine DNA-Bindedomäne vom Typ »$Zn(II)_2Cys_6$«. Von homologen Transkriptionsfaktoren mit der gleichen DNA-Bindedomäne, wie zum Beispiel AoiH (AOI-Cluster, Nakazawa u. a. [2012]), MdpE (Monodictyphenone-Cluster, Chiang u. a. [2010]) und AuR (Aurofusarin-Cluster, Malz u. a. [2005]) wird angenommen, dass sie an ähnlichen Sequenzen binden. Die von MEME und CASSIS ermittelten Bindestellensequenzen von AflR, GliZ und AioH stimmen mit den experimentell ermittelten exakt überein. Die zu AuR ermittelte Bindestelle weist eine starke Ähnlichkeit zu den Bindestellen der homologen Transkriptionsfaktoren auf. Nur die zu MdpE ermittelte Bindestelle hat keine Ähnlichkeit mit den bekannten Sequenzen. Bei dieser Bindestelle könnte ein Fehler von MEME oder CASSIS vorliegen. Allerdings ist dies nur eine Spekulation, da homologe Transkriptionsfaktoren nicht automatisch an die gleiche DNA-Sequenz binden. Für eine Überprüfung der mutmaßlichen Bindestellen fehlt es zur Zeit noch an ausreichend experimentellen Daten [Bulyk 2004; Badis u. a. 2009; Weirauch u. a. 2013].

[10] Der *E-Wert* von MEME gibt die statistische Signifikanz eines Motivs an. Umso kleiner der E-Wert, desto signifikanter ist das Motiv. Der E-Wert ist eine Abschätzung der Anzahl an vergleichbaren Motiven, die in einer gleich großen Menge an zufällig generierten gleich langen Eingabesequenzen gefunden werden (http://meme-suite.org/doc/meme.html, August 2015).

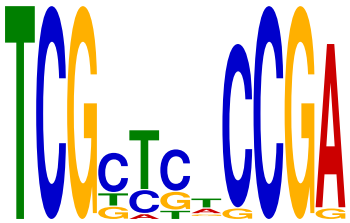

Abbildung 3.1: Logodarstellung der AflR-Bindestelle `TCG` N_5 `CGA` in den Promotoren des Aflatoxin-Gen-Clusters von *Aspergillus flavus*. Erzeugt mit MEME [Bailey u. a. 2015].

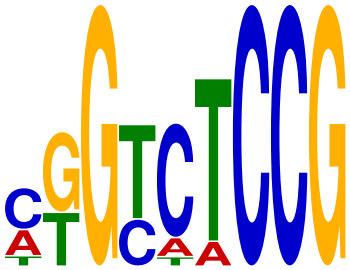

Abbildung 3.2: Logodarstellung der GliZ-Bindestelle `CGG` N_3 `CCG` in den Promotoren des Gliotoxin-Gen-Clusters von *Aspergillus fumigatus*. Erzeugt mit MEME [Bailey u. a. 2015].

3.4 CASSIS

CASSIS wurde entwickelt, um einen deutlichen Nachteil der ähnlichkeitsbasierten SMGC-Vorhersagemethoden (Abschnitt 1.5) auszugleichen: Sie vernachlässigen die Hypothese der Koregulation der Cluster-Gene und die damit verbundenen Informationen. CASSIS, als motivbasierte Methode, beruht hauptsächlich auf der Hypothese der Kolokalisation und Koregulation (Abschnitt 1.7). Die Vorteile dieser alternativen Herangehensweise werden ausführlich in Manuskript 1 (Abschnitt 2.2) und Manuskript 3 (Abschnitt 2.4) diskutiert. In den folgenden Unterabschnitten wird auf die Unterschiede zwischen CASSIS und dessen Vorgänger MDM eingegangen, sowie Besonderheiten bei der Verarbeitung der Eingabedaten durch CASSIS diskutiert.

3.4.1 Unterschiede zwischen CASSIS und MDM

CASSIS ist die Weiterentwicklung von MDM. Beide Programme unterscheiden sich in dreierlei Hinsicht: Zum einen wurden für CASSIS zahlreiche kleinere Verbesserungen vorgenommen, die eher technischer Natur sind. Zum Beispiel wurde die Möglichkeit

der parallelen Ausführung der bis zu 250 MEME- und FIMO-Aufrufe ergänzt. Zum anderen versucht CASSIS verstärkt »unplausible« Motive auszuschließen. Zum Beispiel werden Motive, die sehr oft im gesamten Genom oder gar nicht im Promotor des Ankergens vorkommen, als mögliche cluster-spezifische TFBSs ausgeschlossen.

Einer der wichtigsten Unterschiede ist, wie MDM und CASSIS die Position und Länge eines Gen-Clusters bestimmen. Nach dem »Scannen« nach Motiven lässt MDM ein Fenster (»frame«) mit variabler Länge über die gesamte Genomsequenz gleiten (Abschnitt 2.2). Dabei berechnet MDM für jede Fensterposition die Dichte (»frame score«) der enthaltenen Motivfunde. Das Fenster mit der höchsten Dichte wird als SMGC-Vorhersage ausgegeben, wenn es das Ankergen enthält. Dabei spielt jeder Motivfund eine Rolle für die berechnete Dichte. CASSIS hingegen beginnt die SMGC-Vorhersage am Ankergen und erweitert das Cluster in beide Richtungen, solange sich Promotoren mit mindestens einem Motivfund anschließen (Abschnitt 2.4). Kurze Lücken sind dabei erlaubt, jedoch spielt die Anzahl der Motivfunde pro Promotor eine geringere Rolle: Es wird nur noch zwischen »kein Motivfund« und »mehr als ein Motivfund« unterschieden. Das zweite Verfahren ist einfacher und schneller. Die Cluster-Vorhersagen beider Verfahren unterscheiden sich nicht oder nur geringfügig – je nach betrachtetem SMGC. Nach dem Prinzip »so kompliziert wie nötig und so einfach wie möglich« wurde daher das von MDM genutzte Verfahren aufgegeben. Außerdem ist das von CASSIS genutzte Verfahren weniger anfällig für falsch positive Motivfunde, da die genaue Anzahl an Funden pro Promotor nicht mehr für die Bestimmung der Cluster-Grenzen genutzt wird.

Der praktische Geschwindigkeitsgewinn ist jedoch sehr gering. Er liegt pro vorhergesagtem SMGC im Sekundenbereich. Im Gegensatz: Da CASSIS deutlich mehr Promotoren in der Umgebung des Ankergens für die *de novo* Motivsuche an MEME übergibt (Unterabschnitt 3.3.1), ist die Gesamtlaufzeit deutlich gestiegen. Gleichzeitig besteht bei CASSIS die zuvor angesprochene Möglichkeit, die Laufzeit durch eine massive Parallelisierung der MEME- und FIMO-Aufrufe zu verringern. Manuskript 3 (Abschnitt 2.4, »2.3 SMIPS tool« und »2.5.6 Runtime analysis«) enthält allgemeine Angaben zur Laufzeit von SMIPS und CASSIS.

Wie bereits erwähnt, startet die Cluster-Vorhersage von CASSIS immer bei der Position des Ankergens. Diese feste Bindung an das Ankergen ist gleichzeitig ein Nachteil des Verfahrens. Mit MDM konnten sehr einfach Genomabschnitte identifiziert werden, die sich zwar nicht in der Nähe des Ankergens befanden, aber dennoch eine

hohe Dichte an Motivfunden aufwiesen. So war es möglich, »Sub-Cluster« oder gänzlich neue SMGCs zu entdecken, die möglicherweise komplett oder teilweise vom gleichen Transkriptionsfaktor reguliert werden [Bergmann u. a. 2010]. Mit CASSIS ist dies nicht mehr möglich, da nur noch die direkte Umgebung des Ankergens untersucht wird.

3.4.2 Anwendung von CASSIS auf Genome mit Operons

CASSIS ist nicht für die Vorhersage von SMGCs mit Operons geeignet. Operons besitzen nur einen gemeinsamen Promotor für alle Gene im Operon. CASSIS erwartet aber, dass jedes Gen eine Promotorregion besitzt. Gene ohne Promotorregion können von CASSIS nicht verarbeitet werden (Unterabschnitt 3.4.3). Zukünftige Versionen von CASSIS können um die Option zur Verarbeitung von Genomen mit Operons erweitert werden. Wenn nicht Gene sondern Operons als kleinste Einheit betrachtet werden, können Cluster von koregulierten und kolokalisierten Operons vorhergesagt werden.

Für SMIPS hingegen spielt die Genomorganisation keine Rolle. Es kann in allen prokaryotischen und eukaryotischen Genomen für die Vorhersage von Ankergenen eingesetzt werden, wie Manuskript 3 (Vorhersage in Pilzen), Manuskript 4 (Vorhersage in Bakterien) und Manuskript 5 (ebenfalls Vorhersage in Bakterien) zeigen.

3.4.3 Einfluss der Genomsequenz und Genomannotation auf die Sekundärmetabolit-Gen-Cluster-Vorhersage

Um zu verstehen, warum die Genomsequenz Einfluss auf die Vorhersage von Gen-Clustern hat, soll kurz der Vorgang der Genomassemblierung (»*de novo* genome assembly«) beschrieben werden: Bei der Sequenzierung eines Genoms entstehen sehr viele kurze DNA-Fragmente. Je nach verwendeter Technik zum Beispiel $3 \cdot 10^7$ bis $3 \cdot 10^9$ Fragmente für das menschliche Genom. Anhand von Überlappungen werden die Fragmente von Computerprogrammen wieder zu längeren Sequenzen (»contigs«, »scaffolds«) zusammengesetzt. Dabei entstehen oft Lücken zwischen den Contigs, die ohne zusätzliche Informationen nicht geschlossen werden können [Baker 2012]. Ziel der Genomassemblierung ist die Erzeugung von möglichst wenigen und möglichst langen Contigs.

Die Genauigkeit der von CASSIS vorhergesagten SMGCs hängt unter anderem von der Anzahl der Contigs im assemblierten Genom ab. Bei einer hohen Contig-Anzahl steigt die Wahrscheinlichkeit, dass SMGCs auf mehrere Contigs aufgeteilt sind. Während des in Manuskript 3 durchgeführten Vergleichs zwischen CASSIS, SMURF und antiSMASH wurde festgestellt, dass antiSMASH die Contig-Grenzen ignoriert. Vor dem Start der Ankergen- und SMGC-Vorhersage verknüpft es alle Contigs zu einer einzigen langen DNA-Sequenz. Die Reihenfolge der Contigs entspricht dabei der Reihenfolge in der Eingabedatei. Da die tatsächliche Reihenfolge und Orientierung der Contigs meist nicht bekannt sind, erzeugt dieses Vorgehen künstliche Sequenzen an den Übergängen zwischen den Contigs. CASSIS betrachtet mehrere Contigs zwar ebenfalls als eine zusammenhängende Sequenz, beachtet aber die Contig-Grenzen und ergänzt die Ausgabe um entsprechende Warnhinweise, falls sich eine Vorhersage über mehrere Contigs erstreckt. In diesem Fall versucht CASSIS zusätzlich alternative SMGC-Vorhersagen zu finden, welche die Contig-Grenze nicht überschreiten. Diese alternativen Vorhersagen basieren auf weniger signifikanten Motiven, die ebenfalls in der Umgebung des Ankergens angereichert sind (Abschnitt 1.7).

Weiterhin hängt die Genauigkeit der SMGC-Vorhersagen von der Komplexität der Genomsequenz ab. In Sequenzen niedriger Komplexität (viele Wiederholungen) können die *de novo* vorhergesagten Motive, deren Sequenz dann ebenfalls meist wenig komplex ist, beim »Scannen« (Unterabschnitt 1.6.1) auch an vielen anderen Stellen im Genom gefunden werden. Solche Motive werden von CASSIS direkt verworfen, weil sie deutlich öfter im Genom vorkommen, als bei einer TFBS zu erwarten ist. Im Extremfall werden so alle vorhergesagten Motive verworfen und es können, trotz vorhandenen Ankergenen, keine SMGCs vorhergesagt werden. Dies ist ein Nachteil der motivbasierten SMGC-Vorhersage, verursacht durch die Beschaffenheit der Eingabesequenzen.

In seltenen Fällen kommt es vor, dass die Annotationen von zwei benachbarten Genen überlappen. So zum Beispiel erstreckt sich im Genom von *Aspergillus flavus*[11] das Gen »AFL2G_03678« von Basenpaar 2.133.696 bis 2.134.091 auf dem Minusstrang. Das nachfolgende Gen »AFL2G_03679« beginnt allerdings bereits bei Basenpaar 2.133.927 auf dem Plusstrang. Aus algorithmischer Sicht existiert zwischen den beiden Genen keine intergenische Region und somit auch kein Promotor. CASSIS

[11] Genomsequenz und -annotation: »Aspergillus Comparative Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)«, August 2015

kann diese beiden Gene nicht in die Cluster-Vorhersage einbeziehen, da ohne Promotorsequenz kein Motiv vorhergesagt werden kann. Aus biologischer Sicht gibt es für dieses Phänomen verschiedene Erklärungsmöglichkeiten: Zum einen könnte bei der Sequenzierung des Genoms in dieser Region ein Fehler aufgetreten sein. Im lebenden Organismus wäre die Sequenz dann eine andere und die Position der Gene könnte verschieden sein. Zum anderen könnte eines der beiden Gene falsch annotiert worden sein, also entweder gar nicht existieren oder kürzer sein als angegeben. Schließlich könnte die Annotation auch korrekt sein, beide Gene überlappen sich tatsächlich, und die beteiligten Transkriptionsfaktoren binden innerhalb der kodierenden Bereiche des jeweils anderen Gens [Gerads und Ernst 1998; Quesada u. a. 1999; Kakihara u. a. 2003].

In allen für diese Arbeit verwendeten Genomen (Unterabschnitt 2.4.1, Tabelle S1) kommen 0–10 Überlappungen pro Genom vor. Die einzige Ausnahme bildet *Fusarium graminearum* mit ca. 500 Überlappungen. Für die motivbasierte SMGC-Vorhersage stellen solche Überlappungen kein Problem dar, solange sie nicht innerhalb von vorhergesagten Clustern vorkommen. In allen 38 getesteten SMGCs war dies nicht der Fall. CASSIS gibt eine entsprechende Warnung aus, falls eine Cluster-Vorhersage eine Überlappung enthält. Dann sollte das vorhergesagte SMGC besonders kritisch betrachtet werden. Der »fehlende« Promotor könnte zum Beispiel ausschlaggebend für eine andere Cluster-Grenze sein, wenn er keine Bindestelle eines möglichen cluster-spezifischen Transkriptionsfaktors enthält.

3.5 Nachteile der ähnlichkeitsbasierten Sekundärmetabolit-Gen-Cluster-Vorhersage

Wie in Abschnitt 1.7 bereits erwähnt, ist die motivbasierte SMGC-Vorhersage als Alternative oder Ergänzung zur ähnlichkeitsbasierten SMGC-Vorhersage anzusehen. Mit SMURF, und vor allem antiSMASH, ist die ähnlichkeitsbasierte Vorhersage der zur Zeit dominierende Ansatz. Manuskript 1 (Abschnitt 2.2) und Manuskript 3 (Abschnitt 2.4) diskutieren die Nachteile dieses Ansatzes. Zum Beispiel sind sie auf Vorwissen von bereits bekannten Clustern angewiesen und von eben diesen sind nur wenige bekannt. Außerdem können sie meist nicht zwischen nahe beieinander liegenden Clustern unterscheiden. Schlussendlich ignorieren sie alle Informationen

bezüglich der Koregulation von SMGCs. Insbesondere Manuskript 3 beschreibt einen ausführlichen Vergleich zwischen CASSIS, SMURF und antiSMASH. In Bezug auf diesen Vergleich soll an dieser Stelle auf zwei Eigenheiten von SMURF und antiSMASH aufmerksam gemacht werden, die in den Manuskripten keine direkte Erwähnung finden:

Laut der Publikation von Khaldi u. a. [2010], welche SMURF vorstellt, wurde der Algorithmus ausschließlich mit Daten von 22 bekannten *Aspergillus fumigatus* Clustern trainiert. Dies bedeutet, dass SMURF nur Proteindomänen kennt, die häufig in den SMGCs von *Aspergillus fumigatus* vorkommen. Eine präzise Vorhersage von SMGCs in beliebigen anderen Genomen erscheint mit einem so trainierten Algorithmus eher unwahrscheinlich, da die Zusammensetzung von SMGCs äußerst variabel ist (Abschnitt 1.2).

Laut Scharf u. a. [2011] beinhaltet das Gliotoxin-Gen-Cluster[12] von *Aspergillus fumigatus*[13] die Gene »Afu6g09630« (*gliZ*) bis »Afu6g09745« (*gliH*). Im Zuge des in Manuskript 3 durchgeführten Vergleiches lautete die Vorhersage von antiSMASH »Afu6g09520 bis Afu6g09745«. Die zum Stand August 2015 neueste Version von antiSMASH gibt »Afu6g09520 bis Afu6g09770« aus. SMURF gibt »Afu6g09580 bis Afu6g09740« aus. CASSIS gibt »Afu6g09630 bis Afu6g09785« aus. Medema u. a. [2011] behaupten, das Gliotoxin-Cluster würde von antiSMASH korrekt vorhergesagt beziehungsweise wiedererkannt werden (siehe »Supplementary Table IV«). Tatsächlich stimmt keine der Vorhersagen exakt mit den experimentellen Ergebnissen überein. Die Vorhersage von antiSMASH 2.0 [Blin u. a. 2013] weicht um 11 Gene ab, die Vorhersage von antiSMASH 3.0 [Weber u. a. 2015] um 13 Gene, die Vorhersage von SMURF um 6 Gene und die Vorhersage von CASSIS um 4 Gene. Dieses Beispiel zeigt, dass die ähnlichkeitsbasierten SMGC-Vorhersagemethoden dazu tendieren, Cluster deutlich größer anzugeben, als sie tatsächlich sind.

[12] *Gliotoxin* ist ein Mykotoxin (Schimmelpilzgift), das hauptsächlich vom Pilz *Aspergillus fumigatus* produziert wird. Es hemmt das Immunsystem und wirkt als Zellgift [Scharf u. a. 2011].

[13] Genomsequenz und -annotation: Cerqueira u. a. [2014], »Aspergillus Genome Database« (http://www.aspergillusgenome.org/), März 2014

3.6 Ungeeignete Merkmale für die genaue Vorhersage von Sekundärmetabolit-Gen-Clustern

Zusätzlich zu den Proteindomänen der ähnlichkeitsbasierten Methoden und den TFBSs der motivbasierten Methoden gibt es noch weitere Merkmale, die für die Vorhersage von SMGCs genutzt werden können. Dieser Abschnitt diskutiert Merkmale von SMGCs, die für die Ergänzung von MDM und CASSIS in Betracht gezogen wurden, sich aber als ungeeignet herausstellten.

3.6.1 Genexpression

Andersen u. a. [2013] zeigten, dass Genexpressionsdaten für die Vorhersage von SMGCs genutzt werden können. Mit der Verwendung von Genexpressionsdaten sind aber auch Probleme verbunden: Erstens stehen aufgrund der zusätzlich notwendigen Experimente weniger Expressionsdaten als sequenzierte und annotierte Genome für die Vorhersage von SMGCs zur Verfügung. Zweitens werden die meisten SMGCs unter »normalen« Laborbedingungen nicht exprimiert. Die genauen Bedingungen für die Aktivierung der Cluster sind oft nicht bekannt [Brakhage und Schroeckh 2011]. Demzufolge stehen für diese SMGCs auch keine Expressionsdaten zur Verfügung. Und drittens werden nicht alle Gene eines SMGCs gleichzeitig exprimiert. Die Gene des Aflatoxin-Gen-Clusters zum Beispiel können in früh, mittel und spät exprimierte Gene unterteilt werden [Schmidt-Heydt u. a. 2009; Georgianna und Payne 2009]. Da die Genexpression zu einem bestimmten Zeitpunkt und nicht über einen kontinuierlichen Zeitabschnitt hinweg gemessen wird, besteht die Gefahr, dass zu diesem Zeitpunkt nur ein Teil des SMGCs aktiv war. Die Verwendung solcher Genexpressionsdaten für die SMGC-Vorhersage würde zu ungenauen Cluster-Grenzen führen.

Auf der anderen Seite können durch die Analyse der Genexpression »Sub-Cluster« identifiziert (Unterabschnitt 3.4.1) und SMGCs ohne Ankergene als Startpunkt vorhergesagt werden [Andersen u. a. 2013], wie es mit CASSIS derzeit nicht möglich ist.

3.6.2 DNA-Krümmung und GC-Gehalt

Do und Miyano [2008] untersuchten den GC-Gehalt der DNA von bekannten SMGCs

und deren Umgebung. Abrupte Änderungen im GC-Gehalt sollten die Cluster-Grenzen markieren. Dies war nur bei einem der 26 untersuchten SMGCs der Fall. Zusätzlich weisen nur 6 von 26 SMGCs einen mehr oder weniger konstanten GC-Gehalt innerhalb des Clusters auf.

Der globale Transkriptionsfaktor LaeA beeinflusst den Status des Chromatins[14] und damit die Expression von Genen auf der epigenetischen Ebene [Hoffmeister und Keller 2007]. Do und Miyano [2008] entdeckten einen Zusammenhang zwischen der Regulierung von SMGCs durch LaeA und der Krümmung der DNA im Bereich der Gen-Cluster. Die DNA-Krümmung hängt von ihrer Sequenz ab und kann mit Programmen wie CURVATURE [Shpigelman u. a. 1993] berechnet werden. 15 von 26 SMGCs unterscheiden sich von der umgebenden DNA durch eine geringere Krümmung.

Weder die Krümmung der DNA noch der GC-Gehalt stellen ein Merkmal dar, welches für eine zuverlässige und genaue Vorhersage von SMGCs und deren Cluster-Grenzen genutzt werden kann. Der GC-Gehalt scheint gänzlich ungeeignet zu sein. Die Krümmung der DNA ist für einen Teil der SMGCs als Indikator geeignet. Jedoch nur, um die grobe Position des Clusters zu bestimmen und nicht die exakten Cluster-Grenzen. Genau das ist aber das Ziel der in dieser Arbeit vorgestellten Methoden zur SMGC-Vorhersage.

3.6.3 Länge der intergenischen Regionen

Im Zuge der Entwicklung von CASSIS wurde ebenfalls getestet, ob die Länge der intergenischen Regionen (»Abstand« zwischen den Genen) als Kriterium für die SMGC-Vorhersage genutzt werden kann. Falls SMGCs durch horizontalen Gentransfer zum Beispiel von Bakterien erworben wurden (Abschnitt 3.1), wäre ein Unterschied in der Länge der intergenischen Regionen zwischen Cluster- und Nicht-Cluster-Genen denkbar. Daher wurde für verschiedene bekannte SMGCs (Unterabschnitt 2.4.1, Tabelle S1) überprüft, ob zum Beispiel die Cluster-Grenzen durch besonders große Abstände markiert sind. Jedoch konnte auch mit dieser Hypothese nur ein Teil der bekannten SMGCs vorhergesagt werden. Zum Beispiele weist das Penizillin-Gen-Cluster

[14] Das *Chromatin* ist ein Komplex aus DNA and assoziierten Proteinen bei Eukaryoten, aus denen die Chromosomen bestehen. Euchromatin ist weniger dicht gepackt und stärker transkriptionell aktiv als Heterochromatin.

in *Aspergillus nidulans*[15] eine ungefähr 2.000 bp und 3.500 bp lange intergenische Region an den Rändern des Clusters auf. Die Gene innerhalb des Clusters haben einen Abstand von ungefähr 850 bp. Der genomweite Durchschnitt liegt bei 1.235 bp. Der Abstand innerhalb des Clusters liegt also deutlich unter dem Durchschnitt und der Abstand an den Rändern des Clusters deutlich über dem Durchschnitt. Jedoch ist das Penizillin-Gen-Cluster eines der wenigen positiven Beispiele, bei denen das SMGC eindeutig und fehlerfrei anhand der Länge der intergenischen Regionen erkennbar ist. So wurde auch dieses Merkmal für die Erweiterung von CASSIS verworfen.

3.7 Zukunft der Sekundärmetabolit-Gen-Cluster-Vorhersage

Die computergestützte Vorhersage von SMGCs trägt einen wichtigen Teil zur Entdeckung neuer Wirkstoffe bei. Die Vorhersagen dienen unter anderem als Arbeitshypothesen für neue Experimente. Wie der Vergleich von CASSIS, SMURF und antiSMASH in Manuskript 3 (Abschnitt 2.4) zeigt, können Ankergene des Sekundärmetabolismus und zugehöre SMGCs zuverlässig von Computerprogrammen vorhergesagt werden. Die Genaugkeit der Vorhersagen kann und sollte jedoch noch weiter verbessert werden.

Eine größere Auswahl an experimentellen Daten wird für das Training der Vorhersageprogramme und die Überprüfung ihrer Ergebnisse dringend benötigt. Sowohl die Anzahl der bekannten TFBSs als auch der bekannten SMGCs ist noch zu gering. Zum Beispiel sind weitere experimentell bestätigte TFBS für die Verbesserung und Validierung der *de novo* Motivsuche erforderlich (Unterabschnitt 3.3.3). Vor allem die ähnlichkeitsbasierte SMGC-Vorhersage würde von mehr bekannten SMGC-typischen Proteindomänen profitieren (Abschnitt 1.5). Außerdem sollte die Genauigkeit von SMGC-Vorhersagen nicht nur anhand von bereits bekannten SMGCs ermittelt werden, sondern es sollten verstärkt auch »neue« SMGC-Vorhersagen im Labor überprüft werden.

Für CASSIS könnte die Anwendung einer diskriminativen *de novo* Motivsuche von Vorteil sein. Mögliche Programme dafür wären zum Beispiel DEME [Redhead

[15] Genomsequenz und -annotation: Cerqueira u. a. [2014], »Aspergillus Genome Database« (http://www.aspergillusgenome.org/), März 2014

und Bailey 2007] oder der »diskriminative Modus« von MEME [Narlikar u. a. 2007; Bailey u. a. 2015]. Bei der diskriminativen Suche werden dem Motivsuchealgorithmus zwei verschiedene Mengen an Sequenzen übergeben. Bei der Anwendung mit CASSIS würde die erste (positive) Menge aus den Promotorsequenzen aus der Umgebung des Ankergens bestehen (Unterabschnitt 3.3.1). Die zweite (negative) Menge sollte aus Sequenzen bestehen, die keine cluster-spezifischen TFBSs enthalten. Das könnten zum Beispiel Promotorsequenzen sein, die einen bestimmten Mindestabstand vom Ankergen, zum Beispiel 50 Gene, aufweisen. In der positiven Menge werden dann Motive gesucht, die dort im Vergleich zur negativen Menge angereichert sind. Auf diese Weise können Motive mit biologischer Funktion von zufällig auftretenden Mustern besser unterschieden werden [Redhead und Bailey 2007].

Am aussichtsreichsten für die Verbesserung der SMGC-Vorhersagen erscheint eine Kombination von ähnlichkeits- und motivbasierten Vorhersagemethoden. Auf diese Weise könnten sowohl Informationen auf der Ebene der Proteinfunktion (Proteindomänen) als auch Informationen auf der Ebene der Genregulation (TFBSs) kombiniert genutzt werden. Der Vergleich von Manuskript 3 (Unterabschnitt 2.4.1, Tabelle S3) zeigt deutlich: antiSMASH weist eine höhere Sensitivität als CASSIS auf, dafür ist die Präzision von CASSIS durchgehend besser. Durch eine Kombination der beiden Methoden könnten sowohl Sensitivität als auch Präzision der SMGC-Vorhersage verbessert werden.

Literaturverzeichnis

Amaike und Keller 2011

Amaike, Saori ; Keller, Nancy P.: *Aspergillus flavus.* In: VanAlfen, NK (Hrsg.) ; Bruening, G (Hrsg.) ; Leach, JE (Hrsg.): *Annual Review of Phytopathology, Vol 49* Bd. 49. Palo Alto : Annual Reviews, 2011. – ISBN 978–0–8243–1349–4, S. 107–133. – WOS:000294828400007

Andersen u. a. 2013

Andersen, Mikael R. ; Nielsen, Jakob B. ; Klitgaard, Andreas ; Petersen, Lene M. ; Zachariasen, Mia ; Hansen, Tilde J. ; Blicher, Lene H. ; Gotfredsen, Charlotte H. ; Larsen, Thomas O. ; Nielsen, Kristian F. ; Mortensen, Uffe H.: Accurate prediction of secondary metabolite gene clusters in filamentous fungi. In: *Proceedings of the National Academy of Sciences* 110 (2013), Februar, Nr. 1, E99–E107. `http://dx.doi.org/10.1073/pnas.1205532110`. – DOI 10.1073/pnas.1205532110. – ISSN 0027–8424, 1091–6490

Apostolico u. a. 2000

Apostolico, Alberto ; Bock, Mary E. ; Lonardi, Stefano ; Xu, Xuyan: Efficient detection of unusual words. In: *J. COMP. BIOL* 7 (2000), Nr. 1, S. 71–94

Badis u. a. 2009

Badis, Gwenael ; Berger, Michael F. ; Philippakis, Anthony A. ; Talukder, Shaheynoor ; Gehrke, Andrew R. ; Jaeger, Savina A. ; Chan, Esther T. ; Metzler, Genita ; Vedenko, Anastasia ; Chen, Xiaoyu ; Kuznetsov, Hanna ; Wang, Chi-Fong ; Coburn, David ; Newburger, Daniel E. ; Morris, Quaid ; Hughes, Timothy R. ; Bulyk, Martha L.: Diversity and Complexity in DNA Recognition by Transcription Factors. In: *Science* 324 (2009), Juni,

Nr. 5935, 1720–1723. http://dx.doi.org/10.1126/science.1162327. – DOI 10.1126/science.1162327. – ISSN 0036–8075, 1095–9203

Bailey und Elkan 1994

Bailey, T L. ; Elkan, C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 2 (1994), 28–36. http://www.ncbi.nlm.nih.gov/pubmed/7584402. – ISSN 1553–0833

Bailey u. a. 2009

Bailey, Timothy L. ; Boden, Mikael ; Buske, Fabian A. ; Frith, Martin ; Grant, Charles E. ; Clementi, Luca ; Ren, Jingyuan ; Li, Wilfred W. ; Noble, William S.: MEME Suite: tools for motif discovery and searching. In: *Nucleic Acids Research* 37 (2009), Januar, Nr. suppl 2, W202–W208. http://dx.doi.org/10.1093/nar/gkp335. – DOI 10.1093/nar/gkp335. – ISSN 0305–1048, 1362–4962

Bailey und Elkan 1993

Bailey, Timothy L. ; Elkan, Charles: Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. In: *Machine Learning*, 1993, S. 51–80

Bailey u. a. 2015

Bailey, Timothy L. ; Johnson, James ; Grant, Charles E. ; Noble, William S.: The MEME Suite. In: *Nucleic Acids Research* 43 (2015), Januar, Nr. W1, W39–W49. http://dx.doi.org/10.1093/nar/gkv416. – DOI 10.1093/nar/gkv416. – ISSN 0305–1048, 1362–4962

Baker 2012

Baker, Monya: *De novo* genome assembly: what every biologist should know. In: *Nature Methods* 9 (2012), April, Nr. 4, 333–337. http://dx.doi.org/10.1038/nmeth.1935. – DOI 10.1038/nmeth.1935. – ISSN 1548–7091

Barash u. a. 2003

Barash, Yoseph ; Elidan, Gal ; Friedman, Nir ; Kaplan, Tommy: Modeling

Dependencies in protein-DNA Binding Sites. In: *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*. New York, NY, USA : ACM, 2003 (RECOMB '03). – ISBN 978–1–58113–635–7, 28–37

Baum und Petrie 1966

Baum, Leonard E. ; Petrie, Ted: Statistical Inference for Probabilistic Functions of Finite State Markov Chains. In: *The Annals of Mathematical Statistics* 37 (1966), Dezember, Nr. 6, 1554–1563. http://dx.doi.org/10.1214/aoms/1177699147. – DOI 10.1214/aoms/1177699147. – ISSN 0003–4851, 2168–8990

Bennett und Bentley 1989

Bennett, J.W. ; Bentley, Ronald: What's in a Name? – Microbial Secondary Metabolism. Version: 1989. http://www.sciencedirect.com/science/article/pii/S0065216408703162. In: *Advances in Applied Microbiology* Bd. Volume 34. Academic Press, 1989. – ISBN 0065–2164, 1–28

Bergmann u. a. 2010

Bergmann, Sebastian ; Funk, Alexander N. ; Scherlach, Kirstin ; Schroeckh, Volker ; Shelest, Ekaterina ; Horn, Uwe ; Hertweck, Christian ; Brakhage, Axel A.: Activation of a Silent Fungal Polyketide Biosynthesis Pathway through Regulatory Cross Talk with a Cryptic Nonribosomal Peptide Synthetase Gene Cluster. In: *Applied and Environmental Microbiology* 76 (2010), Dezember, Nr. 24, 8143–8149. http://dx.doi.org/10.1128/AEM.00683-10. – DOI 10.1128/AEM.00683–10. – ISSN 0099–2240, 1098–5336

Bergmann u. a. 2007

Bergmann, Sebastian ; Schumann, Julia ; Scherlach, Kirstin ; Lange, Corinna ; Brakhage, Axel A. ; Hertweck, Christian: Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. In: *Nat Chem Biol* 3 (2007), April, Nr. 4, 213–217. http://dx.doi.org/10.1038/nchembio869. – DOI 10.1038/nchembio869. – ISSN 1552–4450

Blin u. a. 2013

Blin, Kai ; Medema, Marnix H. ; Kazempour, Daniyal ; Fischbach, Michael A. ; Breitling, Rainer ; Takano, Eriko ; Weber, Tilmann: antiSMASH

2.0—a versatile platform for genome mining of secondary metabolite producers. In: *Nucleic Acids Research* 41 (2013), Januar, Nr. W1, W204–W212. http://dx.doi.org/10.1093/nar/gkt449. – DOI 10.1093/nar/gkt449. – ISSN 0305–1048, 1362–4962

Blumenthal 1998

Blumenthal, Thomas: Gene clusters and polycistronic transcription in eukaryotes. In: *BioEssays* 20 (1998), Nr. 6, 480–487. http://dx.doi.org/10.1002/(SICI)1521-1878(199806)20:6<480::AID-BIES6>3.0.CO;2-Q. – DOI 10.1002/(SICI)1521–1878(199806)20:6<480::AID–BIES6>3.0.CO;2–Q. – ISSN 1521–1878

Brakhage 2013

Brakhage, Axel A.: Regulation of fungal secondary metabolism. In: *Nature Reviews Microbiology* 11 (2013), Januar, Nr. 1, 21–32. http://dx.doi.org/10.1038/nrmicro2916. – DOI 10.1038/nrmicro2916. – ISSN 1740–1526

Brakhage und Schroeckh 2011

Brakhage, Axel A. ; Schroeckh, Volker: Fungal secondary metabolites – Strategies to activate silent gene clusters. In: *Fungal Genetics and Biology* 48 (2011), Januar, Nr. 1, 15–22. http://dx.doi.org/10.1016/j.fgb.2010.04.004. – DOI 10.1016/j.fgb.2010.04.004. – ISSN 1087–1845

Bulyk 2004

Bulyk, Martha L.: Computational prediction of transcription-factor binding site locations. In: *Genome Biology* 5 (2004), Nr. 1, 201. http://dx.doi.org/10.1186/gb-2003-5-1-201. – DOI 10.1186/gb–2003–5–1–201. – ISSN 1465–6906

Bérdy 2005

Bérdy, János: Bioactive Microbial Metabolites. In: *The Journal of Antibiotics* 58 (2005), Januar, Nr. 1, 1–26. http://dx.doi.org/10.1038/ja.2005.1. – DOI 10.1038/ja.2005.1. – ISSN 0021–8820

Calvo u. a. 2002

Calvo, Ana M. ; Wilson, Richard A. ; Bok, Jin W. ; Keller, Nancy P.: Relationship between Secondary Metabolism and Fungal Development. In: *Microbiology and Molecular Biology Reviews* 66 (2002), Nr. 3,

447 –459. http://dx.doi.org/10.1128/MMBR.66.3.447-459.2002. – DOI 10.1128/MMBR.66.3.447–459.2002

Caspi u. a. 2014

Caspi, Ron ; Altman, Tomer ; Billington, Richard ; Dreher, Kate ; Foerster, Hartmut ; Fulcher, Carol A. ; Holland, Timothy A. ; Keseler, Ingrid M. ; Kothari, Anamika ; Kubo, Aya ; Krummenacker, Markus ; Latendresse, Mario ; Mueller, Lukas A. ; Ong, Quang ; Paley, Suzanne ; Subhraveti, Pallavi ; Weaver, Daniel S. ; Weerasinghe, Deepika ; Zhang, Peifen ; Karp, Peter D.: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. In: *Nucleic Acids Research* 42 (2014), Januar, Nr. D1, D459–D471. http://dx.doi.org/10.1093/nar/gkt1103. – DOI 10.1093/nar/gkt1103. – ISSN 0305–1048, 1362–4962

Cerqueira u. a. 2014

Cerqueira, Gustavo C. ; Arnaud, Martha B. ; Inglis, Diane O. ; Skrzypek, Marek S. ; Binkley, Gail ; Simison, Matt ; Miyasato, Stuart R. ; Binkley, Jonathan ; Orvis, Joshua ; Shah, Prachi ; Wymore, Farrell ; Sherlock, Gavin ; Wortman, Jennifer R.: The *Aspergillus* Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. In: *Nucleic Acids Research* 42 (2014), Januar, Nr. D1, D705–D710. http://dx.doi.org/10.1093/nar/gkt1029. – DOI 10.1093/nar/gkt1029. – ISSN 0305–1048, 1362–4962

Chiang u. a. 2010

Chiang, Yi-Ming ; Szewczyk, Edyta ; Davidson, Ashley D. ; Entwistle, Ruth ; Keller, Nancy P. ; Wang, Clay C. C. ; Oakley, Berl R.: Characterization of the *Aspergillus nidulans* Monodictyphenone Gene Cluster. In: *Applied and Environmental Microbiology* 76 (2010), Januar, Nr. 7, 2067–2074. http://dx.doi.org/10.1128/AEM.02187-09. – DOI 10.1128/AEM.02187–09. – ISSN 0099–2240, 1098–5336

Collas und Dahl 2008

Collas, Philippe ; Dahl, John A.: Chop it, ChIP it, check it: the current status

of chromatin immunoprecipitation. In: *Frontiers in Bioscience: A Journal and Virtual Library* 13 (2008), S. 929–943. – ISSN 1093–9946

Conlon u. a. 2003

Conlon, Erin M. ; Liu, X. S. ; Lieb, Jason D. ; Liu, Jun S.: Integrating regulatory motif discovery and genome-wide expression analysis. In: *Proceedings of the National Academy of Sciences* 100 (2003), März, Nr. 6, 3339–3344. http://dx.doi.org/10.1073/pnas.0630591100. – DOI 10.1073/pnas.0630591100. – ISSN 0027–8424, 1091–6490

Conway und Boddy 2013

Conway, Kyle R. ; Boddy, Christopher N.: ClusterMine360: a database of microbial PKS/NRPS biosynthesis. In: *Nucleic Acids Research* 41 (2013), Januar, Nr. D1, D402–D407. http://dx.doi.org/10.1093/nar/gks993. – DOI 10.1093/nar/gks993. – ISSN 0305–1048, 1362–4962

Cornish-Bowden 1985

Cornish-Bowden, Athel: Nomenclature for incompletely specified bases in nucleic acid sequences: rcommendations 1984. In: *Nucleic Acids Research* 13 (1985), Oktober, Nr. 9, 3021–3030. http://dx.doi.org/10.1093/nar/13.9.3021. – DOI 10.1093/nar/13.9.3021. – ISSN 0305–1048, 1362–4962

Cortes und Vapnik 1995

Cortes, Corinna ; Vapnik, Vladimir: Support-vector networks. In: *Machine Learning* 20 (1995), September, Nr. 3, 273–297. http://dx.doi.org/10.1007/BF00994018. – DOI 10.1007/BF00994018. – ISSN 0885–6125, 1573–0565

Do und Miyano 2008

Do, Jin ; Miyano, Satoru: The GC and window-averaged DNA curvature profile of secondary metabolite gene cluster in *Aspergillus fumigatus* genome. In: *Applied Microbiology and Biotechnology* 80 (2008), Nr. 5, 841–847. http://dx.doi.org/10.1007/s00253-008-1638-4. – DOI 10.1007/s00253–008–1638–4. – ISSN 0175–7598

Eddy 2011

Eddy, Sean R.: Accelerated Profile HMM Searches. In: *PLoS computational*

biology 7 (2011), Oktober, Nr. 10, S. e1002195. http://dx.doi.org/10.1371/journal.pcbi.1002195. – DOI 10.1371/journal.pcbi.1002195. – ISSN 1553–7358

Eggeling u. a. 2014

Eggeling, Ralf ; Gohr, André ; Keilwagen, Jens ; Mohr, Michaela ; Posch, Stefan ; Smith, Andrew D. ; Grosse, Ivo: On the Value of Intra-Motif Dependencies of Human Insulator Protein CTCF. In: *PLoS ONE* 9 (2014), Januar, Nr. 1, e85629. http://dx.doi.org/10.1371/journal.pone.0085629. – DOI 10.1371/journal.pone.0085629

Ehrlich u. a. 1999

Ehrlich, K.C. ; Montalbano, B.G. ; Cary, J.W.: Binding of the C6-zinc cluster protein, AFLR, to the promoters of aflatoxin pathway biosynthesis genes in *Aspergillus parasiticus*. In: *Gene* 230 (1999), April, Nr. 2, 249–257. http://dx.doi.org/10.1016/S0378-1119(99)00075-X. – DOI 10.1016/S0378–1119(99)00075–X. – ISSN 0378–1119

Ehrlich u. a. 2002

Ehrlich, Kenneth C. ; Montalbano, Beverly G. ; Cary, Jeffrey W. ; Cotty, Peter J.: Promoter elements in the aflatoxin pathway polyketide synthase gene. In: *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 1576 (2002), Juni, Nr. 1-2, 171–175. http://dx.doi.org/10.1016/S0167-4781(02)00282-8. – DOI 10.1016/S0167–4781(02)00282–8. – ISSN 0167–4781

Ettwiller u. a. 2007

Ettwiller, Laurence ; Paten, Benedict ; Ramialison, Mirana ; Birney, Ewan ; Wittbrodt, Joachim: Trawler: *de novo* regulatory motif discovery pipeline for chromatin immunoprecipitation. In: *Nature Methods* 4 (2007), Juli, Nr. 7, 563–565. http://dx.doi.org/10.1038/nmeth1061. – DOI 10.1038/nmeth1061. – ISSN 1548–7091

Fazius u. a. 2011

Fazius, Eugen ; Shelest, Vladimir ; Shelest, Ekaterina: SiTaR: a novel tool for transcription factor binding site prediction. In: *Bioinformatics* (2011). http://dx.doi.org/10.1093/bioinformatics/btr492. – DOI 10.1093/bioinformatics/btr492

Fedorova u. a. 2012

Fedorova, Natalie D. ; Moktali, Venkatesh ; Medema, Marnix H.: Bioinformatics Approaches and Software for Detection of Secondary Metabolic Gene Clusters. Version: Januar 2012. http://link.springer.com/protocol/10.1007/978-1-62703-122-6_2. In: Keller, Nancy P. (Hrsg.) ; Turner, Geoffrey (Hrsg.): *Fungal Secondary Metabolism*. Humana Press, Januar 2012 (Methods in Molecular Biology 944). – ISBN 978–1–62703–121–9 978–1–62703–122–6, 23–45

Fleming 1929

Fleming, Alexander: On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of *B. influenzæ*. In: *British journal of experimental pathology* 10 (1929), Juni, Nr. 3, 226–236. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2048009/. – ISSN 0007–1021

Galas u. a. 1985

Galas, D. J. ; Eggert, M. ; Waterman, M. S.: Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from *Escherichia coli*. In: *Journal of Molecular Biology* 186 (1985), November, Nr. 1, S. 117–128. – ISSN 0022–2836

Galas und Schmitz 1978

Galas, David J. ; Schmitz, Albert: DNAase footprinting a simple method for the detection of protein-DNA binding specificity. In: *Nucleic Acids Research* 5 (1978), Januar, Nr. 9, 3157–3170. http://dx.doi.org/10.1093/nar/5.9.3157. – DOI 10.1093/nar/5.9.3157. – ISSN 0305–1048, 1362–4962

Garner und Revzin 1981

Garner, M M. ; Revzin, A: A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. In: *Nucleic Acids Research* 9 (1981), Juli, Nr. 13, S. 3047–3060. – ISSN 0305–1048

Georgianna und Payne 2009

Georgianna, D. R. ; Payne, Gary A.: Genetic regulation of aflatoxin biosynthesis: From gene to genome. In: *Fungal Genetics and Biology* 46 (2009),

Februar, Nr. 2, 113–125. http://dx.doi.org/10.1016/j.fgb.2008.10.011. – DOI 10.1016/j.fgb.2008.10.011. – ISSN 1087–1845

Gerads und Ernst 1998

Gerads, Michaela ; Ernst, Joachim F.: Overlapping coding regions and trancriptional units of two essential chromosomal genes (CCT8, TRP1) in the fungal pathogen *Candida albicans*. In: *Nucleic Acids Research* 26 (1998), Januar, Nr. 22, 5061–5066. http://dx.doi.org/10.1093/nar/26.22.5061. – DOI 10.1093/nar/26.22.5061. – ISSN 0305–1048, 1362–4962

Ghanbarian und Hurst 2015

Ghanbarian, Avazeh T. ; Hurst, Laurence D.: Neighboring Genes Show Correlated Evolution in Gene Expression. In: *Molecular Biology and Evolution* 32 (2015), Juli, Nr. 7, 1748–1766. http://dx.doi.org/10.1093/molbev/msv053. – DOI 10.1093/molbev/msv053. – ISSN 0737–4038

Grant u. a. 2011

Grant, Charles E. ; Bailey, Timothy L. ; Noble, William S.: FIMO: Scanning for occurrences of a given motif. In: *Bioinformatics* (2011), Februar. http://dx.doi.org/10.1093/bioinformatics/btr064. – DOI 10.1093/bioinformatics/btr064. – ISSN 1367–4803, 1460–2059

Grau u. a. 2006

Grau, J. ; Ben-Gal, I. ; Posch, S. ; Grosse, I.: VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees. In: *Nucleic Acids Research* 34 (2006), Juli, Nr. Web Server, W529–W533. http://dx.doi.org/10.1093/nar/gkl212. – DOI 10.1093/nar/gkl212. – ISSN 0305–1048, 1362–4962

Grau u. a. 2013

Grau, Jan ; Posch, Stefan ; Grosse, Ivo ; Keilwagen, Jens: A general approach for discriminative *de novo* motif discovery from high-throughput data. In: *Nucleic Acids Research* 41 (2013), Januar, Nr. 21, e197–e197. http://dx.doi.org/10.1093/nar/gkt831. – DOI 10.1093/nar/gkt831. – ISSN 0305–1048, 1362–4962

Hellman und Fried 2007

Hellman, Lance M. ; Fried, Michael G.: Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. In: *Nature Protocols* 2 (2007), August, Nr. 8, 1849–1861. http://dx.doi.org/10.1038/nprot.2007.249. – DOI 10.1038/nprot.2007.249. – ISSN 1754–2189

Hoffmeister und Keller 2007

Hoffmeister, Dirk ; Keller, Nancy P.: Natural products of filamentous fungi: enzymes, genes, and their regulation. In: *Natural Product Reports* 24 (2007), März, Nr. 2, 393–416. http://dx.doi.org/10.1039/B603084J. – DOI 10.1039/B603084J. – ISSN 1460–4752

Hopwood 1997

Hopwood, David A.: Genetic Contributions to Understanding Polyketide Synthases. In: *Chem. Rev.* 97 (1997), Nr. 7, 2465–2498. http://dx.doi.org/10.1021/cr960034i. – DOI 10.1021/cr960034i. – ISSN 0009–2665

Hu u. a. 2005

Hu, Jianjun ; Li, Bin ; Kihara, Daisuke: Limitations and potentials of current motif discovery algorithms. In: *Nucleic Acids Research* 33 (2005), Januar, Nr. 15, 4899–4913. http://dx.doi.org/10.1093/nar/gki791. – DOI 10.1093/nar/gki791. – ISSN 0305–1048, 1362–4962

Hughes u. a. 2000

Hughes, J. D. ; Estep, P. W. ; Tavazoie, S. ; Church, G. M.: Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. In: *Journal of Molecular Biology* 296 (2000), März, Nr. 5, S. 1205–1214. http://dx.doi.org/10.1006/jmbi.2000.3519. – DOI 10.1006/jmbi.2000.3519. – ISSN 0022–2836

Hurst u. a. 2004

Hurst, Laurence D. ; Pál, Csaba ; Lercher, Martin J.: The evolutionary dynamics of eukaryotic gene order. In: *Nature Reviews Genetics* 5 (2004), April, Nr. 4, 299–310. http://dx.doi.org/10.1038/nrg1319. – DOI 10.1038/nrg1319. – ISSN 1471–0056

Ichikawa u. a. 2013

Ichikawa, Natsuko ; Sasagawa, Machi ; Yamamoto, Mika ; Komaki, Hisayuki ; Yoshida, Yumi ; Yamazaki, Shuji ; Fujita, Nobuyuki: DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. In: *Nucleic Acids Research* 41 (2013), Januar, Nr. D1, D408–D414. http://dx.doi.org/10.1093/nar/gks1177. – DOI 10.1093/nar/gks1177. – ISSN 0305–1048, 1362–4962

Inglis u. a. 2013

Inglis, Diane O. ; Binkley, Jonathan ; Skrzypek, Marek S. ; Arnaud, Martha B. ; Cerqueira, Gustavo C. ; Shah, Prachi ; Wymore, Farrell ; Wortman, Jennifer R. ; Sherlock, Gavin: Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of *Aspergillus nidulans*, *A. fumigatus*, *A. niger* and *A. oryzae*. In: *BMC Microbiology* 13 (2013), April, Nr. 1, 91. http://dx.doi.org/10.1186/1471-2180-13-91. – DOI 10.1186/1471–2180–13–91. – ISSN 1471–2180

Jenke-Kodama u. a. 2005

Jenke-Kodama, Holger ; Sandmann, Axel ; Müller, Rolf ; Dittmann, Elke: Evolutionary Implications of Bacterial Polyketide Synthases. In: *Molecular Biology and Evolution* 22 (2005), Oktober, Nr. 10, 2027 –2039. http://dx.doi.org/10.1093/molbev/msi193. – DOI 10.1093/molbev/msi193

Jensen u. a. 2004

Jensen, Shane T. ; Liu, X. S. ; Zhou, Qing ; Liu, Jun S.: Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective. In: *Statistical Science* 19 (2004), Februar, Nr. 1, 188–204. http://dx.doi.org/10.1214/088342304000000107. – DOI 10.1214/088342304000000107. – ISSN 0883–4237, 2168–8745

Jones u. a. 2014

Jones, Philip ; Binns, David ; Chang, Hsin-Yu ; Fraser, Matthew ; Li, Weizhong ; McAnulla, Craig ; McWilliam, Hamish ; Maslen, John ; Mitchell, Alex ; Nuka, Gift ; Pesseat, Sebastien ; Quinn, Antony F. ; Sangrador-Vegas, Amaia ; Scheremetjew, Maxim ; Yong, Siew-Yit ; Lopez, Rodrigo ; Hunter, Sarah: InterProScan 5: genome-scale protein function classification. In: *Bioinformatics* 30 (2014), Januar, Nr. 9, 1236–1240.

http://dx.doi.org/10.1093/bioinformatics/btu031. – DOI 10.1093/bioinformatics/btu031. – ISSN 1367–4803, 1460–2059

Kakihara u. a. 2003

Kakihara, Yoshito ; Nabeshima, Kentaro ; Hirata, Aiko ; Nojima, Hiroshi: Overlapping *omt1+* and *omt2+* genes are required for spore wall maturation in *Schizosaccharomyces pombe*. In: *Genes to Cells* 8 (2003), Nr. 6, 547–558. http://dx.doi.org/10.1046/j.1365-2443.2003.00654.x. – DOI 10.1046/j.1365–2443.2003.00654.x. – ISSN 1365–2443

Kanehisa und Goto 2000

Kanehisa, Minoru ; Goto, Susumu: KEGG: Kyoto Encyclopedia of Genes and Genomes. In: *Nucleic Acids Research* 28 (2000), Januar, Nr. 1, 27–30. http://dx.doi.org/10.1093/nar/28.1.27. – DOI 10.1093/nar/28.1.27. – ISSN 0305–1048, 1362–4962

Keller und Hohn 1997

Keller, N P. ; Hohn, T M.: Metabolic pathway gene clusters in filamentous fungi. In: *Fungal Genetics and Biology: FG & B* 21 (1997), Februar, Nr. 1, 17–29. http://www.ncbi.nlm.nih.gov/pubmed/9126615. – ISSN 1087–1845

Keller u. a. 2005

Keller, Nancy P. ; Turner, Geoffrey ; Bennett, Joan W.: Fungal secondary metabolism – from biochemistry to genomics. In: *Nature Reviews Microbiology* 3 (2005), Dezember, Nr. 12, 937–947. http://dx.doi.org/10.1038/nrmicro1286. – DOI 10.1038/nrmicro1286. – ISSN 1740–1526

Khaldi u. a. 2010

Khaldi, Nora ; Seifuddin, Fayaz T. ; Turner, Geoff ; Haft, Daniel ; Nierman, William C. ; Wolfe, Kenneth H. ; Fedorova, Natalie D.: SMURF: Genomic mapping of fungal secondary metabolite clusters. In: *Fungal Genetics and Biology* 47 (2010), September, Nr. 9, 736–741. http://dx.doi.org/10.1016/j.fgb.2010.06.003. – DOI 10.1016/j.fgb.2010.06.003. – ISSN 1087–1845

Kim u. a. 2008

Kim, Nak-Kyeong ; Tharakaraman, Kannan ; Mariño-Ramírez, Leonardo

; Spouge, John L.: Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. In: *BMC Bioinformatics* 9 (2008), Juni, 262. http://dx.doi.org/10.1186/1471-2105-9-262. – DOI 10.1186/1471–2105–9–262. – ISSN 1471–2105

Klepper und Drabløs 2013

Klepper, Kjetil ; Drabløs, Finn: MotifLab: a tools and data integration workbench for motif discovery and regulatory sequence analysis. In: *BMC Bioinformatics* 14 (2013), Januar, Nr. 1, 9. http://dx.doi.org/10.1186/1471-2105-14-9. – DOI 10.1186/1471–2105–14–9. – ISSN 1471–2105

Klepper u. a. 2008

Klepper, Kjetil ; Sandve, Geir K. ; Abul, Osman ; Johansen, Jostein ; Drablos, Finn: Assessment of composite motif discovery methods. In: *BMC Bioinformatics* 9 (2008), Februar, Nr. 1, 123. http://dx.doi.org/10.1186/1471-2105-9-123. – DOI 10.1186/1471–2105–9–123. – ISSN 1471–2105

Korn u. a. 1977

Korn, L J. ; Queen, C L. ; Wegman, M N.: Computer analysis of nucleic acid regulatory sequences. In: *Proceedings of the National Academy of Sciences of the United States of America* 74 (1977), Oktober, Nr. 10, 4401–4405. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431950/. – ISSN 0027–8424

Kulakovskiy u. a. 2010

Kulakovskiy, I. V. ; Boeva, V. A. ; Favorov, A. V. ; Makeev, V. J.: Deep and wide digging for binding motifs in ChIP-Seq data. In: *Bioinformatics* 26 (2010), Oktober, Nr. 20, 2622–2623. http://dx.doi.org/10.1093/bioinformatics/btq488. – DOI 10.1093/bioinformatics/btq488. – ISSN 1367–4803, 1460–2059

Kuttippurathu u. a. 2011

Kuttippurathu, Lakshmi ; Hsing, Michael ; Liu, Yongchao ; Schmidt, Bertil ; Maskell, Douglas L. ; Lee, Kyungjoon ; He, Aibin ; Pu, William T. ; Kong, Sek W.: CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. In: *Bioinformatics* 27 (2011), Januar, Nr. 5, 715–717. http://dx.doi.org/10.1093/bioinformatics/btq707. – DOI 10.1093/bioinformatics/btq707. – ISSN 1367–4803, 1460–2059

Lajoie u. a. 2012

Lajoie, Mathieu ; Gascuel, Olivier ; Lefort, Vincent ; Brehelin, Laurent: Computational discovery of regulatory elements in a continuous expression space. In: *Genome Biology* 13 (2012), November, Nr. 11, R109. `http://dx.doi.org/10.1186/gb-2012-13-11-r109`. – DOI 10.1186/gb–2012–13–11–r109. – ISSN 1465–6906

Lawrence u. a. 1993

Lawrence, C. E. ; Altschul, S. F. ; Boguski, M. S. ; Liu, J. S. ; Neuwald, A. F. ; Wootton, J. C.: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. In: *Science (New York, N.Y.)* 262 (1993), Oktober, Nr. 5131, S. 208–214. – ISSN 0036–8075

Lawrence und Reilly 1990

Lawrence, Charles E. ; Reilly, Andrew A.: An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. In: *Proteins: Structure, Function, and Bioinformatics* 7 (1990), Januar, Nr. 1, 41–51. `http://dx.doi.org/10.1002/prot.340070105`. – DOI 10.1002/prot.340070105. – ISSN 1097–0134

Lawrence 1999

Lawrence, Jeffrey: Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. In: *Current Opinion in Genetics & Development* 9 (1999), Dezember, Nr. 6, 642–648. `http://dx.doi.org/10.1016/S0959-437X(99)00025-8`. – DOI 10.1016/S0959–437X(99)00025–8. – ISSN 0959–437X

Lee und Sonnhammer 2003

Lee, Jennifer M. ; Sonnhammer, Erik L. L.: Genomic Gene Clustering Analysis of Pathways in Eukaryotes. In: *Genome Research* 13 (2003), Januar, Nr. 5, 875–882. `http://dx.doi.org/10.1101/gr.737703`. – DOI 10.1101/gr.737703. – ISSN 1088–9051, 1549–5469

Lemay u. a. 2012

Lemay, Danielle G. ; Martin, William F. ; Hinrichs, Angie S. ; Rijnkels, Monique ; German, J B. ; Korf, Ian ; Pollard, Katherine S.: G-NEST: a

gene neighborhood scoring tool to identify co-conserved, co-expressed genes. In: *BMC Bioinformatics* 13 (2012), September, 253. http://dx.doi.org/10.1186/1471-2105-13-253. – DOI 10.1186/1471–2105–13–253. – ISSN 1471–2105

Li u. a. 2009

Li, Michael H. ; Ung, Peter M. ; Zajkowski, James ; Garneau-Tsodikova, Sylvie ; Sherman, David H.: Automated genome mining for natural products. In: *BMC Bioinformatics* 10 (2009), Juni, Nr. 1, 185. http://dx.doi.org/10.1186/1471-2105-10-185. – DOI 10.1186/1471–2105–10–185. – ISSN 1471–2105

Liao und Zhang 2008

Liao, Ben-Yang ; Zhang, Jianzhi: Coexpression of Linked Genes in Mammalian Genomes Is Generally Disadvantageous. In: *Molecular Biology and Evolution* 25 (2008), August, Nr. 8, 1555–1565. http://dx.doi.org/10.1093/molbev/msn101. – DOI 10.1093/molbev/msn101. – ISSN 0737–4038, 1537–1719

Liu u. a. 2001

Liu, X. ; Brutlag, D. L. ; Liu, J. S.: BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2001), S. 127–138. – ISSN 2335–6936

Lockhart und Winzeler 2000

Lockhart, David J. ; Winzeler, Elizabeth A.: Genomics, gene expression and DNA arrays. In: *Nature* 405 (2000), Juni, Nr. 6788, 827–836. http://dx.doi.org/10.1038/35015701. – DOI 10.1038/35015701. – ISSN 0028–0836

Ma u. a. 2012

Ma, Xiaotu ; Kulkarni, Ashwinikumar ; Zhang, Zhihua ; Xuan, Zhenyu ; Serfling, Robert ; Zhang, Michael Q.: A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. In: *Nucleic Acids Research* 40 (2012), Januar, Nr. 7, e50–e50. http://dx.doi.org/10.1093/nar/gkr1135. – DOI 10.1093/nar/gkr1135. – ISSN 0305–1048, 1362–4962

Malz u. a. 2005

Malz, Sascha ; Grell, Morten N. ; Thrane, Charlotte ; Maier, Frank J.

; Rosager, Pernille ; Felk, Angelika ; Albertsen, Klaus S. ; Salomon, Siegfried ; Bohn, Lisbeth ; Schäfer, Wilhelm ; Giese, Henriette: Identification of a gene cluster responsible for the biosynthesis of aurofusarin in the *Fusarium graminearum* species complex. In: *Fungal Genetics and Biology* 42 (2005), Mai, Nr. 5, 420–433. http://dx.doi.org/10.1016/j.fgb.2005.01.010. – DOI 10.1016/j.fgb.2005.01.010. – ISSN 1087–1845

Matys u. a. 2003

Matys, V. ; Fricke, E. ; Geffers, R. ; Gössling, E. ; Haubrock, M. ; Hehl, R. ; Hornischer, K. ; Karas, D. ; Kel, A. E. ; Kel-Margoulis, O. V. ; Kloos, D.-U. ; Land, S. ; Lewicki-Potapov, B. ; Michael, H. ; Münch, R. ; Reuter, I. ; Rotert, S. ; Saxel, H. ; Scheer, M. ; Thiele, S. ; Wingender, E.: TRANSFAC®: transcriptional regulation, from patterns to profiles. In: *Nucleic Acids Research* 31 (2003), Januar, Nr. 1, 374 –378. http://dx.doi.org/10.1093/nar/gkg108. – DOI 10.1093/nar/gkg108

McGary u. a. 2013

McGary, Kriston L. ; Slot, Jason C. ; Rokas, Antonis: Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. In: *Proceedings of the National Academy of Sciences of the United States of America* 110 (2013), Juli, Nr. 28, 11481–11486. http://dx.doi.org/10.1073/pnas.1304461110. – DOI 10.1073/pnas.1304461110. – ISSN 0027–8424

Medema u. a. 2011

Medema, Marnix H. ; Blin, Kai ; Cimermancic, Peter ; Jager, Victor d. ; Zakrzewski, Piotr ; Fischbach, Michael A. ; Weber, Tilmann ; Takano, Eriko ; Breitling, Rainer: antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. In: *Nucleic Acids Research* 39 (2011), Januar, Nr. suppl 2, W339–W346. http://dx.doi.org/10.1093/nar/gkr466. – DOI 10.1093/nar/gkr466. – ISSN 0305–1048, 1362–4962

Michalak 2008

Michalak, Pawel: Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. In: *Genomics* 91 (2008), März, Nr.

3, 243–248. http://dx.doi.org/10.1016/j.ygeno.2007.11.002. – DOI 10.1016/j.ygeno.2007.11.002. – ISSN 0888–7543

Mootz u. a. 2002

Mootz, Henning D. ; Schwarzer, Dirk ; Marahiel, Mohamed A.: Ways of assembling complex natural products on modular nonribosomal peptide synthetases. In: *Chembiochem: A European Journal of Chemical Biology* 3 (2002), Juni, Nr. 6, 490–504. http://dx.doi.org/10.1002/1439-7633(20020603)3:6<490::AID-CBIC490>3.0.CO;2-N. – DOI 10.1002/1439–7633(20020603)3:6<490::AID–CBIC490>3.0.CO;2–N. – ISSN 1439–4227

Nakazawa u. a. 2012

Nakazawa, Takehito ; Ishiuchi, Kan'ichiro ; Praseuth, Alex ; Noguchi, Hiroshi ; Hotta, Kinya ; Watanabe, Kenji: Overexpressing Transcriptional Regulator in *Aspergillus oryzae* Activates a Silent Biosynthetic Pathway to Produce a Novel Polyketide. In: *ChemBioChem* 13 (2012), April, Nr. 6, 855–861. http://dx.doi.org/10.1002/cbic.201200107. – DOI 10.1002/cbic.201200107. – ISSN 1439–7633

Narlikar u. a. 2007

Narlikar, Leelavati ; Gordân, Raluca ; Hartemink, Alexander J.: Nucleosome Occupancy Information Improves *de novo* Motif Discovery. Version: 2007. http://link.springer.com/chapter/10.1007/978-3-540-71681-5_8. In: Speed, Terry (Hrsg.) ; Huang, Haiyan (Hrsg.): *Research in Computational Molecular Biology*. Springer Berlin Heidelberg, 2007 (Lecture Notes in Computer Science 4453). – ISBN 978–3–540–71680–8 978–3–540–71681–5, 107–121

Neuwald u. a. 1995

Neuwald, Andrew F. ; Liu, Jun S. ; Lawrence, Charles E.: Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. In: *Protein Science* 4 (1995), August, Nr. 8, 1618–1632. http://dx.doi.org/10.1002/pro.5560040820. – DOI 10.1002/pro.5560040820. – ISSN 1469–896X

Newman und Cragg 2012

Newman, David J. ; Cragg, Gordon M.: Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. In: *Journal of Natural Products* 75

(2012), März, Nr. 3, 311–335. http://dx.doi.org/10.1021/np200906s. – DOI 10.1021/np200906s. – ISSN 0163–3864

Nguyen und Androulakis 2009

Nguyen, Tung T. ; Androulakis, Ioannis P.: Recent Advances in the Computational Discovery of Transcription Factor Binding Sites. In: *Algorithms* 2 (2009), März, Nr. 1, 582–605. http://dx.doi.org/10.3390/a2010582. – DOI 10.3390/a2010582

Pavesi u. a. 2004

Pavesi, Giulio ; Mauri, Giancarlo ; Pesole, Graziano: *In silico* representation and discovery of transcription factor binding sites. In: *Briefings in Bioinformatics* 5 (2004), Januar, Nr. 3, 217–236. http://dx.doi.org/10.1093/bib/5.3.217. – DOI 10.1093/bib/5.3.217. – ISSN 1467–5463, 1477–4054

Pavesi u. a. 2006

Pavesi, Giulio ; Mereghetti, Paolo ; Zambelli, Federico ; Stefani, Marco ; Mauri, Giancarlo ; Pesole, Graziano: MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. In: *Nucleic Acids Research* 34 (2006), Juli, Nr. Web Server issue, W566–W570. http://dx.doi.org/10.1093/nar/gkl285. – DOI 10.1093/nar/gkl285. – ISSN 0305–1048

Pearl 1984

Pearl, Judea: *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc., 1984. – ISBN 978–0–201–05594–8

Posch u. a. 2007

Posch, Stefan ; Grau, Jan ; Gohr, Andre ; Ben-Gal, Irad ; Kel, Alexander E. ; Grosse, Ivo: Recognition of cis-regulatory elements with VOMBAT. In: *Journal of Bioinformatics and Computational Biology* 5 (2007), April, Nr. 2B, S. 561–577. – ISSN 0219–7200

Quesada u. a. 1999

Quesada, Víctor ; Ponce, María R. ; Micol, José L.: OTC and AUL1, two

convergent and overlapping genes in the nuclear genome of *Arabidopsis thaliana*. In: *FEBS Letters* 461 (1999), November, Nr. 1–2, 101–106. http://dx.doi.org/10.1016/S0014-5793(99)01426-X. – DOI 10.1016/S0014–5793(99)01426–X. – ISSN 0014–5793

Rausch u. a. 2005

Rausch, Christian ; Weber, Tilmann ; Kohlbacher, Oliver ; Wohlleben, Wolfgang ; Huson, Daniel H.: Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). In: *Nucleic Acids Research* 33 (2005), Januar, Nr. 18, 5799–5808. http://dx.doi.org/10.1093/nar/gki885. – DOI 10.1093/nar/gki885. – ISSN 0305–1048, 1362–4962

Redhead und Bailey 2007

Redhead, Emma ; Bailey, Timothy L.: Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. In: *BMC Bioinformatics* 8 (2007), Oktober, Nr. 1, 385. http://dx.doi.org/10.1186/1471-2105-8-385. – DOI 10.1186/1471–2105–8–385. – ISSN 1471–2105

Rosenberg und Court 1979

Rosenberg, M ; Court, D: Regulatory Sequences Involved in the Promotion and Termination of RNA Transcription. In: *Annual Review of Genetics* 13 (1979), Nr. 1, 319–353. http://dx.doi.org/10.1146/annurev.ge.13.120179.001535. – DOI 10.1146/annurev.ge.13.120179.001535

Roven und Bussemaker 2003

Roven, Crispin ; Bussemaker, Harmen J.: REDUCE: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. In: *Nucleic Acids Research* 31 (2003), Januar, Nr. 13, 3487–3490. http://dx.doi.org/10.1093/nar/gkg630. – DOI 10.1093/nar/gkg630. – ISSN 0305–1048, 1362–4962

Rubin und Green 2013

Rubin, Alan F. ; Green, Phil: Expression-based segmentation of the Drosophila genome. In: *BMC Genomics* 14 (2013), November, Nr. 1, 812. http://dx.doi.

org/10.1186/1471-2164-14-812. – DOI 10.1186/1471–2164–14–812. – ISSN 1471–2164

Röttig u. a. 2011

Röttig, Marc ; Medema, Marnix H. ; Blin, Kai ; Weber, Tilmann ; Rausch, Christian ; Kohlbacher, Oliver: NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. In: *Nucleic Acids Research* 39 (2011), Januar, Nr. suppl 2, W362–W367. http://dx.doi.org/10.1093/nar/gkr323. – DOI 10.1093/nar/gkr323. – ISSN 0305–1048, 1362–4962

Sandve u. a. 2007

Sandve, Geir ; Abul, Osman ; Walseng, Vegard ; Drabløs, Finn: Improved benchmarks for computational motif discovery. In: *BMC Bioinformatics* 8 (2007), Juni, Nr. 1, 193. http://dx.doi.org/10.1186/1471-2105-8-193. – DOI 10.1186/1471–2105–8–193. – ISSN 1471–2105

Sandve und Drabløs 2006

Sandve, Geir K. ; Drabløs, Finn: A survey of motif discovery methods in an integrated framework. In: *Biology Direct* 1 (2006), April, 11. http://dx.doi.org/10.1186/1745-6150-1-11. – DOI 10.1186/1745–6150–1–11. – ISSN 1745–6150

Scharf u. a. 2011

Scharf, Daniel H. ; Heinekamp, Thorsten ; Remme, Nicole ; Hortschansky, Peter ; Brakhage, Axel A. ; Hertweck, Christian: Biosynthesis and function of gliotoxin in *Aspergillus fumigatus*. In: *Applied Microbiology and Biotechnology* 93 (2011), November, Nr. 2, 467–472. http://dx.doi.org/10.1007/s00253-011-3689-1. – DOI 10.1007/s00253–011–3689–1. – ISSN 0175–7598, 1432–0614

Schmidt-Heydt u. a. 2009

Schmidt-Heydt, Markus ; Abdel-Hadi, Ahmed ; Magan, Naresh ; Geisen, Rolf: Complex regulation of the aflatoxin biosynthesis gene cluster of *Aspergillus flavus* in relation to various combinations of water activity and temperature. In: *International Journal of Food Microbiology* 135 (2009), November, Nr. 3,

231–237. http://dx.doi.org/10.1016/j.ijfoodmicro.2009.07.026. – DOI 10.1016/j.ijfoodmicro.2009.07.026. – ISSN 0168–1605

Schroeckh u. a. 2009

Schroeckh, Volker ; Scherlach, Kirstin ; Nützmann, Hans-Wilhelm ; Shelest, Ekaterina ; Schmidt-Heck, Wolfgang ; Schuemann, Julia ; Martin, Karin ; Hertweck, Christian ; Brakhage, Axel A.: Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in *Aspergillus nidulans*. In: *Proceedings of the National Academy of Sciences* 106 (2009), August, Nr. 34, 14558–14563. http://dx.doi.org/10.1073/pnas.0901870106. – DOI 10.1073/pnas.0901870106. – ISSN 0027–8424, 1091–6490

Shpigelman u. a. 1993

Shpigelman, E. S. ; Trifonov, E. N. ; Bolshoy, A.: CURVATURE: software for the analysis of curved DNA. In: *Computer applications in the biosciences : CABIOS* 9 (1993), Januar, Nr. 4, 435–440. http://dx.doi.org/10.1093/bioinformatics/9.4.435. – DOI 10.1093/bioinformatics/9.4.435. – ISSN 1367–4803, 1460–2059

Simcha u. a. 2012

Simcha, David ; Price, Nathan D. ; Geman, Donald: The Limits of *De Novo* DNA Motif Discovery. In: *PLoS ONE* 7 (2012), November, Nr. 11, e47836. http://dx.doi.org/10.1371/journal.pone.0047836. – DOI 10.1371/journal.pone.0047836

Sinha und Tompa 2003a

Sinha, S. ; Tompa, M.: Performance comparison of algorithms for finding transcription factor binding sites. In: *Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings*, 2003, S. 214–220

Sinha und Tompa 2003b

Sinha, Saurabh ; Tompa, Martin: YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. In: *Nucleic Acids Research* 31 (2003), Januar, Nr. 13, 3586–3588. http://dx.doi.org/10.1093/nar/gkg618. – DOI 10.1093/nar/gkg618. – ISSN 0305–1048, 1362–4962

Spellberg u. a. 2008

Spellberg, Brad ; Guidos, Robert ; Gilbert, David ; Bradley, John ; Boucher, Helen W. ; Scheld, W. M. ; Bartlett, John G. ; Edwards, John ; America, the Infectious Diseases Society o.: The Epidemic of Antibiotic-Resistant Infections: A Call to Action for the Medical Community from the Infectious Diseases Society of America. In: *Clinical Infectious Diseases* 46 (2008), Januar, Nr. 2, 155–164. http://dx.doi.org/10.1086/524891. – DOI 10.1086/524891. – ISSN 1058–4838, 1537–6591

Starcevic u. a. 2008

Starcevic, Antonio ; Zucko, Jurica ; Simunkovic, Jurica ; Long, Paul F. ; Cullum, John ; Hranueli, Daslav: ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. In: *Nucleic Acids Research* 36 (2008), Dezember, Nr. 21, 6882–6892. http://dx.doi.org/10.1093/nar/gkn685. – DOI 10.1093/nar/gkn685. – ISSN 0305–1048

Stormo 2000

Stormo, G. D.: DNA binding sites: representation and discovery. In: *Bioinformatics (Oxford, England)* 16 (2000), Januar, Nr. 1, S. 16–23. – ISSN 1367–4803

Stormo u. a. 1982

Stormo, G. D. ; Schneider, T. D. ; Gold, L. ; Ehrenfeucht, A.: Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. In: *Nucleic Acids Research* 10 (1982), Mai, Nr. 9, S. 2997–3011. – ISSN 0305–1048

Tae u. a. 2009

Tae, Hongseok ; Sohng, Jae K. ; Park, Kiejung: Development of an analysis program of type I polyketide synthase gene clusters using homology search and profile hidden Markov model. In: *Journal of Microbiology and Biotechnology* 19 (2009), Februar, Nr. 2, S. 140–146. – ISSN 1017–7825

Takeda u. a. 2014

Takeda, Itaru ; Umemura, Myco ; Koike, Hideaki ; Asai, Kiyoshi ; Machida, Masayuki: Motif-Independent Prediction of a Secondary Metabolism Gene Cluster Using Comparative Genomics: Application to Sequenced Genomes of

Aspergillus and Ten Other Filamentous Fungal Species. In: *DNA Research* (2014), April, dsu010. http://dx.doi.org/10.1093/dnares/dsu010. – DOI 10.1093/dnares/dsu010. – ISSN 1340–2838, 1756–1663

Thijs u. a. 2002

Thijs, Gert ; Marchal, Kathleen ; Lescot, Magali ; Rombauts, Stephane ; De Moor, Bart ; Rouzé, Pierre ; Moreau, Yves: A Gibbs Sampling Method to Detect Overrepresented Motifs in the Upstream Regions of Coexpressed Genes. In: *Journal of Computational Biology* 9 (2002), April, Nr. 2, 447–464. http://dx.doi.org/10.1089/10665270252935566. – DOI 10.1089/10665270252935566. – ISSN 1066–5277

Thompson u. a. 2007

Thompson, William A. ; Newberg, Lee A. ; Conlan, Sean ; McCue, Lee A. ; Lawrence, Charles E.: The Gibbs Centroid Sampler. In: *Nucleic Acids Research* 35 (2007), Juli, S. W232–W237. http://dx.doi.org/10.1093/nar/gkm265. – DOI 10.1093/nar/gkm265. – ISSN 0305–1048. – WOS:000255311500044

Tompa u. a. 2005

Tompa, Martin ; Li, Nan ; Bailey, Timothy L. ; Church, George M. ; Moor, Bart D. ; Eskin, Eleazar ; Favorov, Alexander V. ; Frith, Martin C. ; Fu, Yutao ; Kent, W. J. ; Makeev, Vsevolod J. ; Mironov, Andrei A. ; Noble, William S. ; Pavesi, Giulio ; Pesole, Graziano ; Régnier, Mireille ; Simonis, Nicolas ; Sinha, Saurabh ; Thijs, Gert ; Helden, Jacques v. ; Vandenbogaert, Mathias ; Weng, Zhiping ; Workman, Christopher ; Ye, Chun ; Zhu, Zhou: Assessing computational tools for the discovery of transcription factor binding sites. In: *Nature Biotechnology* 23 (2005), Nr. 1, 137–144. http://dx.doi.org/10.1038/nbt1053. – DOI 10.1038/nbt1053. – ISSN 1087–0156

Umemura u. a. 2013

Umemura, Myco ; Koike, Hideaki ; Nagano, Nozomi ; Ishii, Tomoko ; Kawano, Jin ; Yamane, Noriko ; Kozone, Ikuko ; Horimoto, Katsuhisa ; Shin-ya, Kazuo ; Asai, Kiyoshi ; Yu, Jiujiang ; Bennett, Joan W. ; Machida, Masayuki: MIDDAS-M: Motif-Independent *De Novo* Detection of Secondary Metabolite Gene Clusters through the Integration of Genome Sequencing and Transcriptome Data. In: *PLoS ONE* 8 (2013), Dezember, Nr. 12, e84028.

http://dx.doi.org/10.1371/journal.pone.0084028. – DOI 10.1371/journal.pone.0084028

VéZina u. a. 1975

VÉZina, Claude ; Kudelski, Alicia ; Sehgal, S. N.: Rapamycin (AY-22,989), a new antifungal antibiotic. I. Taxonomy of the producing streptomycete and isolation of the active principle. In: *The Journal of Antibiotics* 28 (1975), Nr. 10, 721–726. http://dx.doi.org/10.7164/antibiotics.28.721. – DOI 10.7164/antibiotics.28.721. – ISSN 0021–8820, 1881–1469

Walker u. a. 2012

Walker, Michael B. ; King, Benjamin L. ; Paigen, Kenneth: Clusters of Ancestrally Related Genes That Show Paralogy in Whole or in Part Are a Major Feature of the Genomes of Humans and Other Species. In: *PLoS ONE* 7 (2012), April, Nr. 4, e35274. http://dx.doi.org/10.1371/journal.pone.0035274. – DOI 10.1371/journal.pone.0035274

Weber u. a. 2009

Weber, T. ; Rausch, C. ; Lopez, P. ; Hoof, I. ; Gaykova, V. ; Huson, D.H. ; Wohlleben, W.: CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. In: *Journal of Biotechnology* 140 (2009), März, Nr. 1-2, 13–17. http://dx.doi.org/10.1016/j.jbiotec.2009.01.007. – DOI 10.1016/j.jbiotec.2009.01.007. – ISSN 01681656

Weber 2014

Weber, Tilmann: *In silico* tools for the analysis of antibiotic biosynthetic pathways. In: *International Journal of Medical Microbiology* 304 (2014), Mai, Nr. 3–4, 230–235. http://dx.doi.org/10.1016/j.ijmm.2014.02.001. – DOI 10.1016/j.ijmm.2014.02.001. – ISSN 1438–4221

Weber u. a. 2015

Weber, Tilmann ; Blin, Kai ; Duddela, Srikanth ; Krug, Daniel ; Kim, Hyun U. ; Bruccoleri, Robert ; Lee, Sang Y. ; Fischbach, Michael A. ; Müller, Rolf ; Wohlleben, Wolfgang ; Breitling, Rainer ; Takano, Eriko ; Medema, Marnix H.: antiSMASH 3.0—a comprehensive resource for the genome

mining of biosynthetic gene clusters. In: *Nucleic Acids Research* (2015), Mai, gkv437. http://dx.doi.org/10.1093/nar/gkv437. – DOI 10.1093/nar/gkv437. – ISSN 0305–1048, 1362–4962

Weirauch u. a. 2013

Weirauch, Matthew T. ; Cote, Atina ; Norel, Raquel ; Annala, Matti ; Zhao, Yue ; Riley, Todd R. ; Saez-Rodriguez, Julio ; Cokelaer, Thomas ; Vedenko, Anastasia ; Talukder, Shaheynoor ; Dream5 Consortium ; Bussemaker, Harmen J. ; Morris, Quaid D. ; Bulyk, Martha L. ; Stolovitzky, Gustavo ; Hughes, Timothy R.: Evaluation of methods for modeling transcription factor sequence specificity. In: *Nature Biotechnology* 31 (2013), Februar, Nr. 2, 126–134. http://dx.doi.org/10.1038/nbt.2486. – DOI 10.1038/nbt.2486. – ISSN 1087–0156

Wray u. a. 2003

Wray, Gregory A. ; Hahn, Matthew W. ; Abouheif, Ehab ; Balhoff, James P. ; Pizer, Margaret ; Rockman, Matthew V. ; Romano, Laura A.: The Evolution of Transcriptional Regulation in Eukaryotes. In: *Molecular Biology and Evolution* 20 (2003), Januar, Nr. 9, 1377–1419. http://dx.doi.org/10.1093/molbev/msg140. – DOI 10.1093/molbev/msg140. – ISSN 0737–4038, 1537–1719

Yadav u. a. 2003

Yadav, Gitanjali ; Gokhale, Rajesh S. ; Mohanty, Debasisa: SEARCHPKS: a program for detection and analysis of polyketide synthase domains. In: *Nucleic Acids Research* 31 (2003), Januar, Nr. 13, 3654–3658. http://dx.doi.org/10.1093/nar/gkg607. – DOI 10.1093/nar/gkg607. – ISSN 0305–1048, 1362–4962

Yamamoto und Casida 1999

Yamamoto, Izuru (Hrsg.) ; Casida, John E. (Hrsg.): *Nicotinoid Insecticides and the Nicotinic Acetylcholine Receptor*. Tokyo : Springer Japan, 1999 http://link.springer.com/10.1007/978-4-431-67933-2. – ISBN 978–4–431–68011–6 978–4–431–67933–2

Yi u. a. 2007

Yi, Gangman ; Sze, Sing-Hoi ; Thon, Michael R.: Identifying clusters of functionally related genes in genomes. In: *Bioinformatics* 23 (2007), Januar, Nr.

9, 1053–1060. http://dx.doi.org/10.1093/bioinformatics/btl673. – DOI 10.1093/bioinformatics/btl673. – ISSN 1367–4803, 1460–2059

Yu u. a. 2004

Yu, Jiujiang ; Chang, Perng-Kuang ; Ehrlich, Kenneth C. ; Cary, Jeffrey W. ; Bhatnagar, Deepak ; Cleveland, Thomas E. ; Payne, Gary A. ; Linz, John E. ; Woloshuk, Charles P. ; Bennett, Joan W.: Clustered Pathway Genes in Aflatoxin Biosynthesis. In: *Applied and Environmental Microbiology* 70 (2004), März, Nr. 3, 1253 –1262. http://dx.doi.org/10.1128/AEM.70.3.1253-1262.2004. – DOI 10.1128/AEM.70.3.1253–1262.2004

Zambelli u. a. 2013

Zambelli, Federico ; Pesole, Graziano ; Pavesi, Giulio: Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. In: *Briefings in Bioinformatics* 14 (2013), Januar, Nr. 2, 225–237. http://dx.doi.org/10.1093/bib/bbs016. – DOI 10.1093/bib/bbs016. – ISSN 1467–5463, 1477–4054

Abbildungsverzeichnis

1.1 Schema eines Sekundärmetabolit-Gen-Clusters 20
1.2 Beispiel für ein Sekundärmetabolit-Gen-Cluster 20
1.3 Schema verschiedener PKS-Typen 22
1.4 Schema verschiedener NRPS-Typen 23
1.5 Beispiel für eine Konsensussequenz 28
1.6 Beispiel für ein (Sequenz-)Profil 28

3.1 Logodarstellung der AflR-Bindestelle TCG N_5 CGA 120
3.2 Logodarstellung der GliZ-Bindestelle CGG N_3 CCG 120

Ehrenwörtliche Erklärung

1 Die geltende Promotionsordnung der biologisch-pharmazeutischen Fakultät ist mir bekannt.

2 Die vorliegende Dissertation habe ich selbst angefertigt. Ich habe keine Textabschnitte eines Dritten oder eigene Prüfungsarbeiten ohne Kennzeichnung übernommen. Ich habe alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen angegeben.

3 Unterstützung bei der Auswahl und Auswertung des Materials, sowie bei der Herstellung der Manuskripte, habe ich nur von den genannten Koautoren und in der Danksagung genannten Personen erhalten.

4 Die Hilfe eines Promotionsberaters habe ich nicht in Anspruch genommen. Dritte haben weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

5 Ich habe die Dissertation nicht bereits zuvor als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht.

6 Ich habe die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei keiner anderen Hochschule als Dissertation eingereicht.

Jena, 12. Juli 2017